USING DEEP LEARNING TO AUTOMATICALLY EXTRACT PSYCHOLOGICAL REPRESENTATIONS OF COMPLEX NATURAL STIMULI

Craig Aaron Sanders

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the Department of Psychological and Brain Sciences
Indiana University
July, 2018

Accepted by the Graduate Faculty, Indiana University, in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Doctoral Committee

_____

Robert Nosofsky, PhD

_____

Robert Goldstone, PhD

_____

Chen Yu, PhD

_____

David Crandall, PhD

July 12, 2018

Acknowledgments

I would first like to thank everyone on my dissertation committee for their input on this work: Robert Goldstone, Chen Yu, David Crandall, and of course my advisor, Robert Nosofsky. Rob, I have learned so much working with you, and I am grateful that you always continued to support me, no matter how many programming mistakes I made. And I have to admit that I never thought when I applied to grad school that I would end up writing a dissertation about rocks, but I think we've managed to create a really cool data set together, and I look forward to analyzing it for years to come.

To all my family and friends back in Michigan, especially my parents Edmund Sanders and Anita Sanders, I would like to thank you always for believing in me, even if you didn't always understand what I was doing. I initially wanted to move as far away as I could for grad school, but I'm glad now that I ultimately stayed within driving distance.

Adrian Paneto, Brian Meagher, Cory Chew, Edward Koenig, and Oliver Lees, I would like to thank all of you for making my dissertation year immensely more enjoyable through our weekly D&D adventures. Grad students often feel guilty when they don't think they're being productive, but I've come to appreciate that sometimes taking a break and having fun is the most productive thing you can do.

Finally, I would like to thank my wonderful girlfriend, Natalie Rodriguez Quintana. Natalie, despite all the hardships of grad school, broken bones, and everything else we've been through, you have made these past three years the best of my life. Without your love and support, this dissertation would not have been possible.

Craig Aaron Sanders

USING DEEP LEARNING TO AUTOMATICALLY EXTRACT PSYCHOLOGICAL REPRESENTATIONS OF COMPLEX NATURAL STIMULI

Cognitive psychologists have developed many formal models of categorization, but they have been almost exclusively tested using artificial categories because deriving psychological representations of natural stimuli using traditional methods such as multidimensional scaling (MDS) has been an intractable task. In this dissertation, I show that deep convolutional neural networks (CNNs) may be used to solve this problem. First, I provide an overview of how CNNs work, and I review related work that has examined the relationship between the representations learned by CNNs and the psychological representations used by humans. I then demonstrate that CNNs can be trained to predict the MDS coordinates of rocks derived in previous work (Nosofsky, Sanders, Meagher, & Douglas, 2017). In Experiment 1, I conduct a conceptual replication of Nosofsky et al.'s (2017) methods and demonstrate that similar MDS dimensions emerge across different sets of rocks, and the CNNs are able to generalize from one set to the other. Then in Experiment 2, I conduct a categorization experiment and demonstrate that the CNN representations can be used in conjunction with a formal cognitive model to predict human behavior, indicating that CNNs can be used to automate MDS studies in the future.

_____

Robert Nosofsky, PhD

_____

Robert Goldstone, PhD

_____

Chen Yu, PhD

_____

David Crandall, PhD

Contents

Introduction

Categorization is a fundamental task students in the sciences must learn. Astronomy students must learn to categorize stars, biology students must learn to categorize organisms, and geology students must learn to categorize rocks. Consider the rocks shown in Figure 1 which belong to the categories obsidian, anthracite, and rhyolite. Notice that despite belonging to two different categories, the examples of obsidian and anthracite are all visually similar—they are all black, shiny rocks. On the other hand, notice that despite all belonging to the same category, the examples of rhyolite are all visually dissimilar—they all have different colors and patterns. Making things even more confusing is that that despite being more similar to anthracite, obsidian actually belongs to the same high-level category as rhyolite, which is igneous, whereas anthracite belongs to the high-level category of metamorphic. This high amount of between-category similarity and low amount of within-category similarity makes teaching rocks a challenging task.

Cognitive psychologists have proposed many formal models of categorization that can quantitatively predict people's categorization behavior (see Pothos & Wills, 2011 for a review). These models have the potential to be very useful for educators because they could be used, for example, to determine which examples of categories should be shown, and in what order, to maximize students' accuracy (e.g., Mathy & Feldman, 2016; Patil, Zhu, Kopeć, & Love, 2014). However, before this potential can be realized, it must be demonstrated that formal categorization models can be applied to real-world categories, and thus far they have mostly been tested using artificial stimuli. Common stimuli include Gabor patches (e.g., Maddox, Ashby, & Bohil, 2003), simple shapes with inscribed lines (e.g., Nosofsky, 1986), and dot patterns (e.g., (Posner & Keele, 1968; see Richler & Palmeri, 2014 for other examples). Such stimuli are convenient for modeling purposes because their dimensions are mathematically well-

defined. A Gabor patch, for instance, can be defined in terms of its orientation and spatial frequency, which makes it straightforward to use as input to a formal model.

The dimensions of natural stimuli may not be so apparent or easily quantified, however. Again, consider the rocks in Figure 1. These are complex natural stimuli, and people's psychological representations of them may involve any number of loosely-defined dimensions, such as color, texture, attractiveness, the presence of spots, the presence of stripes, etc. Recently, Nosofsky and colleagues have shown that multidimensional scaling (MDS; Kruskal & Wish, 1978) can be used to extract people's psychological representations of images of rocks[1] (Nosofsky, Sanders, Gerdom, Douglas, & McDaniel, 2017; Nosofsky, Sanders, Meagher, et al., 2017) . MDS will be discussed more formally in a later section of this dissertation, but in brief, it is a technique in which similarity judgments are collected for pairs of stimuli, and then the stimuli are placed in an n-dimensional space such that similar items are close together in the space and dissimilar items are far apart. The resulting dimensions can then be interpreted and used as psychological representations of stimuli for modeling purposes. Nosofsky et al. (2017) conducted MDS analyses using a data set of 360 rocks and found that the derived dimensions had sensible psychological interpretations, such as lightness of color and grain size. Nosofsky et al. found that in many cases rocks from different categories were closer together in this space than rocks of the same category, supporting the intuition that in the domain of rocks, there is often a high amount of between-category similarity and a low amount of within-category similarity, and suggesting that there may be important exceptions to the family-resemblance principle that is generally thought to govern natural categories (Rosch & Mervis, 1975). Moreover, it was found that these MDS dimensions could be used in conjunction with a formal

[1] Nosofsky et al.'s results are based on data collected from people who reported no geological training. Deriving psychological representations of expert geologists is an ongoing research project.

model to predict people's categorization behavior (Nosofsky, Sanders, & McDaniel, 2018; Nosofsky, Sanders, Zhu, & McDaniel, submitted). These results indicate that psychological theory could be used to better understand how students learn categories of rocks and to make recommendations to educators.

Despite these successes, the MDS approach has some significant limitations. For one, MDS cannot generalize to new stimuli. Nosofsky et al.'s MDS solution provides psychological representations of the 360 rocks in their data set, but only those 360, and thus their results will have limited practical applications for geoscience educators, who would be interested in teaching the full range of rocks that exist in the real world. The naïve solution to this limitation would be to simply conduct more MDS analyses using bigger data sets, but this is impractical, if not impossible. Creating an MDS solution with *N* stimuli requires constructing an *NxN* similarity matrix, where each cell indicates the average similarity between two stimuli. This means that tens of thousands of similarity judgments had to be collected just to create the MDS solution for the 360 rock dataset—collecting this much data was actually so time- and resource-prohibitive, that Nosofsky et al.'s MDS solution was ultimately based upon a similarity matrix where most cells were based on only one or two observations, and many cells were left completely empty. Scaling this method to even larger datasets is simply not feasible.

An automated system for extracting psychological representations would be extremely useful because it would allow formal models to be applied to arbitrary stimuli without having to conduct time and resource-intensive similarity-judgment studies. While this goal may have seemed unattainable just a few years ago, it may now be possible using recent advances in the field of deep learning (LeCun, Bengio, & Hinton, 2015). Like traditional connectionist networks, deep learning networks transform data through layers of computational units, with the

connections between units being trained to produce a desired output from the input using the backpropagation algorithm (Rumelhart, Hinton, & Williams, 1986). Unlike traditional connectionist networks, however, which typically only have one hidden layer, deep learning networks may have dozens or even hundreds of layers with, each one producing a new hierarchical representation of the input data. A particular class of deep learning networks known as convolutional neural networks (CNNs) have proven to be powerful tools for extracting information from images and are now used for automating many complex tasks, from handwritten digit recognition (e.g., LeCun et al., 1989), to image classification (e.g., Krizhevsky, Sutskever, & Hinton, 2012), to scene captioning (e.g., Vinyals, Toshev, Bengio, & Erhan, 2015).

In this dissertation I will show that CNNs may also be trained to extract psychological representations from images. Specifically, I will show that CNNs can be trained to extract the MDS coordinates of the rocks from Nosofsky et al.'s (2017) data set. Moreover, I will demonstrate that the trained networks can accurately produce the MDS coordinates of novel rocks, and these representations can be used in conjunction with a formal psychological model to predict human categorization behavior. These findings indicate that CNNs may be used to automatically derive psychological representations for an unlimited number of stimuli, making categorization research using natural categories more feasible.

The rest of this dissertation will be organized as follows: First, I will provide a brief overview of how CNNs function, and I will discuss some related work in cognitive psychology and neuroscience that has explored the relationship between the representations learned by CNNs and the psychological representations used by humans. I will then describe my deep learning procedure: I will provide an overview of Nosofsky et al.'s (2017) data set, and I will describe how several CNN-based models were trained and evaluated on their ability to extract the MDS

coordinates from the images of rocks in this data set. Then in Experiment 1, I conduct MDS analyses on a new set of rocks to evaluate how well the models can predict the MDS dimensions and similarity relationships of novel stimuli. Finally, in Experiment 2 I compare the model-predicted coordinates and the actual MDS coordinates on their ability to predict human categorization behavior when used in conjunction with formal psychological models. Materials and code used in this dissertation may be found online (https://osf.io/d6b9y/).

Overview of Deep Convolutional Neural Networks

To understand CNNs, first consider a multilayer perceptron, as illustrated in Figure 2. This type of network is made entirely of *fully-connected layers*—that is, every unit in a given layer is connected to every unit in the previous layer. The input to a unit is given by the sum of the unit's bias with the dot product of the incoming units' activations and their connection weights. A nonlinear activation is then applied to this input. Traditional connectionist networks typically use sigmoidal functions as their activation functions, but such functions have very small gradients for both large and small inputs, making learning slow. Modern deep learning networks instead make use of rectified linear units (ReLU; Nair & Hinton, 2010). These units use $f(x) = \max(0, x)$ as their activation function, which allows for better gradient propagation throughout the network, resulting in more efficient learning. Nonlinear activation functions allow multilayer perceptrons to construct nonlinear transformations of their input data in their hidden layers, and it has been shown that this allows such networks to approximate any function (Cybenko, 1989). In practice, though, networks made only of fully-connected layers do not scale well to high-dimensional data, such as images. For example, consider a network that takes color images (with red, blue, and green channels) as input. Even if the images only have a modest resolution of 200x200 pixels, each unit in a fully-connected hidden layer would require $200 \times 200 \times 3 = 120{,}000$ connections! A network with such a huge number of parameters would be slow to train and hard to fit in computer memory, and it would also be prone to overfitting to noise and failing to generalize to new data.

CNNs make use of *convolutional layers* and *pooling layers* to process images and other sources of high-dimensional data more efficiently.

Figure 3A illustrates a convolutional layer. This layer takes an input matrix, X, and creates an output matrix, X', using a filter (also called a feature map), which is defined in terms of a set of weights, W, and a bias, b (the values of which are again learned through backpropagation). The size of W is known as the filter's receptive field. X' is formed by computing dot products between W and local regions of X, then adding the bias term. For example, the filter in

Figure 3A has a 3x3 receptive field, and applying this filter to the local region of X indicated by the blue box results in a value of 1 in X', whereas applying this filter to the local region of X indicated by the orange box results in a value of -5 in X' (notice that the spatial arrangement of values in X' corresponds to the spatial arrangement of the local regions in X). Mathematically, this operation is discrete convolution, hence the name "convolutional layer" (LeCun et al., 2015). After performing this convolution operation, the values are typically transformed again using a nonlinear activation function, such as ReLU. While

Figure 3A only displays a single two-dimensional filter due to space constraints, a typical convolutional layer will have many filters, with each set of W and b values learned through backpropagation. And since images are typically treated as three-dimensional input (with length, width, and color channels), filters applied to image data are also typically three-dimensional.

A CNN's filters may be interpreted as neurons that respond to specific types of features. The filter in

Figure 3A, for example, may be interpreted as a vertical edge detector—it only produces a positive output when its input has greater values in the middle column than the left and right columns. One significant advantage of CNNs over classic computer vision methods is that these features are learned from the data and do not need to be engineered by hand. These learned

features form a hierarchical structure. Simple features learned in early layers, such as edges and color splotches, may in later layers be combined into shapes and patterns, and the filters in final layers of a CNN may respond to extremely abstract features such as the presence of white flowers or people in specific postures (Szegedy et al., 2013). Filters are shared for each local region of the space, which not only reduces the number of parameters needed to learn by the model, but also gives the network a degree of location invariance (i.e., the network can detect features no matter where they are located in an image).

Convolutional layers in a CNN are often followed by pooling layers, which reduce the spatial size of their input. The two most common pooling operations are max pooling, which reduces local regions of the input to their maximum value, and average pooling, which reduces local regions of the input to their average value.

Figure 3B shows pooling layers with 2x2 receptive fields; the blue and orange square show specific local regions before and after pooling. While it may seem counterintuitive to insert layers into a network that destroy information, this further reduces the number of parameters the network needs, and reducing the resolution of the input may help the network recognize features by removing noise. Occasionally global pooling layers are also used, which pool across the entire filter. This is useful, for example, for converting 2D or 3D filters into 1D vectors that can then feed into fully-connected layers. Sequences of convolutional and pooling layers are often followed by a number fully-connected layers before the final output layer, although some CNN architectures do not have any fully-connected layers other than the final classification-response layer (e.g., He, Zhang, Ren, & Sun, 2016; Szegedy, Vanhoucke, Ioffe, Shlens, & Wojna, 2016),

and some do not have any pooling layers and are solely made up of convolutional layers (e.g., Springenberg, Dosovitskiy, Brox, & Riedmiller, 2014).

Modern CNNs also take advantage of many computational techniques that have been developed in recent years. A full review of all these techniques goes beyond the scope of this dissertation, but here I will outline some of the major ones that I make use of in the present work. For example, many CNNs make use of dropout (Srivastava, Hinton, Krizhevsky, Sutskever, & Salakhutdinov, 2014), a technique in which some proportion of units in a given layer are removed from the network during each step of training. Concretely, the activation of each unit is set to 0 with some probability (the dropout rate). Essentially this means that during each step of training, a subnetwork is sampled from the full network. During test, all unit activations are kept but are reweighted according to the dropout rate, which effectively averages the output of all subnetworks. Dropout can therefore be conceptualized as improving generalization by training an ensemble of subnetworks. Alternatively, dropout can be conceptualized as a regularization technique. When dropout is used, a hidden unit cannot rely on any one unit from the previous layer being active, so it must learn to distribute weight across all of its input, which again improves the networks' generalization.

Another modern innovation that many CNNs take advantage of is batch normalization (Ioffe & Szegedy, 2015). In this technique, the unit activations of a layer are normalized to have a specific mean and variance (the values of which are learned by backpropagation). This speeds up training because gradient descent is more efficient when parameters are at a similar scale. This also increases generalization because the distribution of activations in a given layer is less sensitive to the distribution of activations in the previous layers, so the network becomes more robust to differences between the distributions of training and test data.

Other modern techniques involve modifying the learning algorithm. In classic backpropagation, the chain rule of derivatives is used to calculate the gradient of each parameter with respect to some error function, such as the mean squared error between the network's output and the desired output. Each parameter is then adjusted by adding to it its negative gradient multiplied by a fixed *learning rate* (generally a small constant). This latter step is a form of gradient descent, and it can be conceptualized as a constant downhill motion across the error space, ending at a local minimum. In *gradient descent with momentum,* each parameter is adjusted proportionally to a running average of the gradients from the last several training batches, rather than just the gradient from the current training batch. The relative weightings of the current gradient and the prior gradients is given by a *momentum* parameter, with higher momentum values placing more weight on the current gradient. This technique effectively smooths out the bumps in the error surface, allowing the network to reach better minima more quickly (Bengio, 2012). Kingman and Ba's "Adam" (2014) is a recently popular optimization algorithm that is similar to gradient descent with momentum, but it additionally scales the learning rate of each parameter according to the magnitude of its gradient, which makes traversing the error space more efficient, further increasing the learning speed.

With this background in mind, in the next section I discuss related work from cognitive psychology and neuroscience that has explored the relationship between the representations learned by deep CNNs and the psychological representations used by humans.

Related Work

Using connectionist networks as a means for extracting psychological representations is actually an old idea. Rumelhart and Todd (1993) showed that a shallow network could be used to extract representations of Morse codes. Their network took pairs of codes as input, and each code was transformed into learned hidden-layer representations, from which similarities were computed using hardcoded measures. They found that after training such a network using similarity judgments collected from humans, the network's hidden layers represented the codes in terms of their length and whether they were made of mostly dots or dashes, which were the same dimensions uncovered through multidimensional scaling (Rothkopf, 1957; Shepard, 1963). Moreover, these representations generalized to stimuli that the network was not trained on.

While conceptually similar to my own approach, Rumelhart and Todd's (1993) differs in that they did not train the network to directly produce MDS dimensions; rather, they had the network indirectly learn psychological representations by training it to produce human similarity judgments. This approach may initially seem more natural since the MDS dimensions are not ground truth values, but are instead derived from the similarity judgments. However, one function of MDS analyses is to remove noise in similarity judgments, and recall that Nosofsky et al.'s MDS solution is based on a noisy, incomplete similarity matrix, so the MDS data may actually be cleaner and more suitable for training networks. Another virtue of my own approach is that it is extensible. There are certain features, such as holes, that are not common enough to significantly influence people's similarity judgments and appear in the MDS solution, but are nonetheless important for modeling categorization (the presence of holes is a strong indicator that a rock is pumice). It is not clear how Rumelhart and Todd's approach would accommodate

such dimensions, but in my approach, it is straightforward to train CNNs to recognize these dimensions by manually adding them to the feature-space representation.

More modern research suggests that the hidden representations learned by deep networks pre-trained to perform computer vision tasks may in some instances correspond to human representations. These networks are most commonly pre-trained to perform image classification on the ILSVRC data set, which consists of over one million images of natural objects belonging to 1000 categories (Russakovsky et al., 2015). Representations of images can be created by providing those images to the networks as input, and then extracting the unit activations from one or more of the networks' hidden layers. It has been found, for instance, that these sorts of representations can be used to predict human fixation durations (Kümmerer, Theis, & Bethge, 2014; Kümmerer, Wallis, & Bethge, 2017). Furthermore, Peterson, Abbot, and Griffiths (2017) found that unlike the representations extracted using classic computer vision algorithms, CNN representations of images of animals could predict human similarity judgments of those images, and preliminary work suggests that these representations may also be able to predict human categorization behavior when used in conjunction with formal psychological models (Battleday, Peterson, & Griffiths, 2017). Similarly, Lake et al. (2015b) found that human typicality ratings for various category exemplars were correlated with the strength of category responses in CNNs, but they were not correlated with responses from models trained using classic computer vision features. There were some interesting discrepancies between the human judgments and CNN responses in this study, however, that might suggest somewhat different forms of representations. For example, humans gave the highest typicality ratings to perfectly yellow bananas, indicating that their ratings were based on how "ideal" the banana was. In contrast, CNNs gave equally high responses to green or spotted bananas, indicating that their responses

were based on how common the features of the bananas actually were (most bananas are not perfectly yellow).

Some researchers have proposed that CNNs may actually be good models of the human visual system. The architecture of CNNs, which build increasingly complex hierarchical representations from lower-level input, is similar to the architectures of models of the visual cortex that have been proposed in the past (e.g., Riesenhuber & Poggio, 1999), lending some credence to this viewpoint. It has also been shown that CNN representations can predict brain activity in the visual cortices of both human and nonhuman primates (Agrawal, Stansbury, Malik, & Gallant, 2014; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014), and there is a strong correlation between the categorization accuracy of CNNs and the amount of neural activity they can explain in the IT cortex (Agrawal et al., 2014; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). It does not seem to be generally true, however, that CNNs with higher categorization accuracy have more human-like representations. Kheradpisheh et al. (2016) found that the representations of CNNs with the highest classification accuracies were not as similar to human representations (derived from category confusion matrices) as some CNNs with lower accuracies.

While these findings provide some preliminary evidence that pre-trained CNNs may provide a ready source of stimulus representations for cognitive modeling, there are still reasons to prefer MDS and other classic models of psychological representations. For instance, CNN representations are notoriously difficult to interpret, while uncovering semantically-interpretable dimensions is one of the principal reasons for conducting an MDS analysis, and this interpretability is important for establishing scientific theory. And while decades of research has shown that MDS dimensions make useful models of human representations, there is reason to be

skeptical of the extent to which CNN representations really correspond to human representations. One such reason is that while CNNs and the human visual system may be similar at a macro scale, their computational details are very different. CNNs take static grids of pixels with a constant spatial resolution as input, but human eyes take in new information with every saccade and have a high resolution in the fovea but a low resolution in the periphery (Akbas & Eckstein, 2017). Furthermore, while typical CNNs have completely feedforward architectures, the human visual system contains many feedback and recurrent connections (Olshausen, 2013), and backpropagation is not regarded as a biologically-plausible learning mechanism because there is no known mechanism for distributing error signals across biological neurons (e.g., Crick, 1989; Stork, 1989).

Another reason to be skeptical of the correspondence between human and CNN representations is that, in many cases, humans and CNNs behave in qualitatively different ways. For instance, deep CNNs gradually learn to recognize visual categories through hundreds or thousands of feedback trials, and while this is similar to how very young infants learn object names, by the age of two humans can learn whole categories from single examples (L. B. Smith, Jayaraman, Clerkin, & Yu, 2018). Human toddlers may have a learning advantage because, unlike CNNs which are passively trained on static images, toddlers actively manipulate their environment to create better training data. In support of this view, it has been shown that CNNs trained to recognize objects from egocentric video data achieve higher accuracy when the data is collected from toddlers rather than adults because toddlers tend to produce more diverse viewpoints of objects, and the objects also tend to occupy a larger portion of the visual field (Bambach, Crandall, Smith, & Yu, 2016; Bambach et al., 2016; Bambach, Zhang, Crandall, & Yu, 2017). Furthermore, Liu et al. (2017) derived the sequence of frames from this egocentric

video data set that would allow CNNs to learn the objects in the fewest number of iterations, and it was found that the optimal sequence was qualitatively similar to the actual visual experience of toddlers (i.e., the optimal sequence consisted of contiguous presentations of the same object, with different viewpoints smoothly transitioning into each other). These results indicate that a better understanding of how humans learn visual objects may lead to better techniques for training computer vision systems.

In a direct comparison between human and CNN behavior, Lake et al. (2015a) found that CNNs were much worse than humans at both recognizing and generating novel symbols after being presented with single examples. In another direct comparison, Eckstein et al. (2017) found that in a visual search task where both humans and CNNs had to report the presence or absence of target items in natural scenes, humans had a higher miss rate than the CNNs when the target was not scaled proportionally with the rest of the scene (e.g., humans were less likely to report the presence of a toothbrush if the toothbrush was the size of a broom). On the other hand, humans also had a lower false alarm rate for superficially-similar distractor items with inappropriate scales (e.g., CNNs were more likely to mistake a hot-air balloon with a soccer-ball pattern for an actual soccer ball). Rajalingham et al. (2018) tested humans, monkeys, and CNNs on their ability to discriminate various categories of images. They found that the CNNs produced similar confusion matrixes to humans and monkeys when the confusion matrixes were aggregated across categories (e.g., on average, humans, monkeys, and CNNs all confused dogs with camels more often than they confused dogs with houses). However, the CNNs' confusions for individual images were significantly different from those observed for humans or monkeys. Similarly, Kheradpisheh et al. (2016) found that humans and CNNs produced similar confusion matrixes for images with uniform backgrounds, but they made very different confusions when

objects were placed in natural backgrounds. And Geirhos et al. (2017) found that while CNNs may match or even exceed human accuracy for some image classification tasks, humans become much more accurate when noise is added to the images, or the images are converted to greyscale. These findings indicate that humans and CNNs may have different representations, or at the very least, they make use of their representations in different ways.

It is also well known that CNNs can be fooled into misclassifying images by adding to the images tiny perturbations imperceptible to the human eye (Szegedy et al., 2013). For example, an image of a school bus that a CNN would normally classify correctly may be subtly modified so that the CNN would instead classify it as an ostrich, even though a human observer may consider the modified image to look identical to the original. Furthermore, images that appear to be nothing more than noise to humans may be confidently classified by CNNs into meaningful categories such as "robin" or "bubble" (Nguyen, Yosinski, & Clune, 2014). The existence of these *adversarial examples* indicates that CNN representations are more brittle than those used by humans, casting further doubt on their adequacy as models of psychological representations.

Furthermore, adversarial examples are surprisingly robust. Networks with different architectures trained using different data sets may be fooled by the same examples; even entirely different classes of machine learning models, such as support vector machines and decision trees, may be fooled by the same adversarial examples (Papernot, McDaniel, & Goodfellow, 2016). Adversarial examples may even manifest in the physical world; an adversarial example can be generated on a computer, printed out, and then digitally photographed again, and the photograph may fool a CNN just as well as the original (Kurakin, Goodfellow, & Bengio, 2016). While some researchers have discovered techniques for making CNNs more robust to adversarial examples

(Goodfellow, Shlens, & Szegedy, 2014; Gu & Rigazio, 2014; Papernot, McDaniel, Wu, Jha, & Swami, 2016), other researchers have also discovered new algorithms for generating them (e.g., Carlini & Wagner, 2016). These results all indicate that there may be fundamental differences between current computer vision systems and the human vision system, and new paradigms will need to emerge before one can be used to understand the other.

As a counterpoint, though, Elsayed et al. (2018) conducted an experiment which indicated that humans may actually be sensitive to some adversarial examples, too. Participants classified images into two different categories, such as cats and dogs, or spiders and snakes. In *image trials*, participants were presented with an unmodified image belonging to one of the two categories. In *adversarial trials*, participants were presented with an image belonging to one category but modified to fool CNNs into classifying the image as the opposite category. In *false trials*, participants were presented with an image that belonged to neither category but was modified to fool CNNs into classifying the image as one of the two. It was found that accuracy was lower in the adversarial trials compared to the image trials. This finding alone would not necessarily suggest that the adversarial examples were more confusable with the opposing category—the modified images could have simply had more perceptual noise. However, participants were also more likely to respond with the adversarially-targeted category in the false trials. In other words, when participants were tasked with classifying images as cats or dogs, and they were presented with a picture of a spider that had been modified so that CNNs would classify it as a dog, the human participants were also more likely to respond with "dog." It should be noted, however, that this this was a modest effect—participants were only about 1%-5% more likely to choose the adversarially-targeted category. Participants were also only able to

view the images for 63-71 ms, so it is not clear that these findings could be replicated or have any real-world significance.

All in all, the current evidence for CNN representations being useful models of human representations seems to be equivocal, so for the present research I do not directly treat CNNs as models of the human visual system. Instead, I treat CNNs as pure machine learning models that can be trained to reproduce MDS dimensions, which are well-established models of human representations. As I will discuss more in the general discussion, direct comparisons between the abilities of CNN, MDS, and other forms of representation to predict human behavior will be a fruitful area of future research.

Deep Learning Procedure

The goal of the deep learning procedure was to train models that could take images of rocks as input and produce their 8-dimensional MDS coordinates as output[2]. Once the models are trained in this way, they can be used to automatically generate psychological representations of infinite numbers of rocks. In this section I describe the data set I used and how that data was split to train and evaluate the models. I then describe the models themselves, identify the best one, and comment on how well it is able to generalize to novel input it was not trained on.

Data Set

In this dissertation I make use of Nosofsky et al.'s (2017) 360 rock data set, which was downloaded from that manuscript's companion website (https://osf.io/w64fv/). This data set consists of 360 images of rocks belonging to the 3 high-level categories of igneous, metamorphic, and sedimentary, with 10 subtypes within each high-level category, and 12 individual tokens within each subtype. The exact subtypes used in this data set can be found in Tables

Table *1*. These subtypes were intended to be representative of those found in introductory geology textbooks. All of the images had a width of 800 pixels, but they varied in their height. The images were scaled to the default input resolutions of each network, with the edges being cropped as necessary to make the images square without distorting their aspect ratios. The data set also contains the values of each image along 8 psychological dimensions, derived using

[2] All of the models reported here produce the 8 MDS dimensions simultaneously. An alternative implementation would be to train 8 different versions of each model, with each one producing only one of the MDS dimensions as output. While rigorous comparisons between these alternatives are not reported here due to the time-intensive nature of training several different versions of the same models, preliminary analysis suggested that models that produced all dimensions simultaneously actually had more accurate predictions. This is in line with previous research on multitask learning which has found that training networks to perform multiple tasks leads to better performance on each individual task (Caruana, 1998).

multidimensional-scaling (MDS). To reiterate from the introduction, MDS is a technique for deriving psychological representations in which similarity judgments are collected for pairs of stimuli, and then the stimuli are placed in an n-dimensional space such that similar items are close together in the space and dissimilar items are far apart.

More concretely, Nosofsky et al. (2017) collected similarity judgments for pairs of rocks in this data set on a 1-9 scale, with 1 being most dissimilar and 9 being most similar. They then used these judgments to construct a 360x360 similarity matrix, where each cell, s_{ij}, indicates the similarity judgment between item *i* and item *j*. As mentioned in the introduction, a 360x360 similarity matrix has tens of thousands of cells. Because of this huge data requirement, many of the values of these cells were based on only 1 or 2 observations, and some cells were actually left empty. This means that there is probably some noise in the MDS solution, and as shall be seen, this may have adversely impacted the deep learning networks' predictive power.

A maximum-likelihood version of MDS (M. D. Lee, 2001) was used to fit x_{im} values to this similarity matrix, where x_{im} is the value of stimulus *i* along MDS dimension *m*. In this MDS solution, the predicted similarities between items ($\hat{s}_{ij}$) are assumed to be a decreasing linear function of the Euclidean distances in the space (d_{ij}):

$$\hat{s}_{ij} = u - v \cdot d_{ij} \qquad (1)$$

where *u*, *v*, and are free parameters and d_{ij} is given by the x_{im} values. Assuming that the observed similarity judgments are Gaussian distributed with variance σ^2 around the predicted similarity judgments, then finding the x_{im} that maximize the likelihood of the observed similarities is equivalent to finding the x_{im} that minimize the sum of squared deviations (SSD) between the predicted ($\hat{s}_{ij}$) and observed (s_{ij}) similarities:

$$SSD = \sum_{i<j} \left(\hat{s}_{ij} - s_{ij}\right)^2 \quad (2)$$

Nosofsky et al. (2017) conducted computer searches to find the best fitting x_{im} values when using 1-12 MDS dimensions[3]. They chose the 8-dimensional solution as the best solution based on model fit diagnostics and the high interpretability and psychological meaningfulness of the resulting dimensions.

The initial dimensions given by an MDS solution may be hard to interpret because the Euclidean metric is rotation-invariant, and thus the orientation of the space is arbitrary. To aid in the interpretation of the MDS dimensions, Nosofsky et al. (2017) also collected direct ratings of the rocks on a 1-9 scale along a set of hypothesized dimensions, such as lightness of color, average grain size, etc. Procrustes analyses (Gower & Dijksterhuis, 2004) were conducted to rotate, translate, and scale the space so that the MDS dimensions would maximally correspond with these dimensions. Formally, let r_{im} be the average direct rating along hypothesized dimension *m*, and let x'_{im} be the value of item *i* along dimension *m* following rotation, translation, and scaling. Computer searches were conducted to find the rotation, translation, and scaling parameters that minimized the sum of squared deviations (SSD) between the x'_{im} and r_{im} values:

$$SSD = \sum_i \sum_m (x'_{im} - r_{im})^2 \quad (3)$$

The final rotated MDS solution was produced by removing the dimension-scaling operation from the x'_{im} values so that the magnitude of the pairwise distances in the un-rotated space would be preserved. Figure 4 - Figure *8* show this rotated MDS space graphically

[3]The initial x_{im} values for these computer searches were given by MATLAB's MDSCALE function, which fits a standard non-metric MDS solution (Kruskal, 1964). I take the same approach when fitting the new MDS solution described later in this article.

(interactive versions of these plots can be viewed online: https://osf.io/w64fv/), with each figure showing the location of all 360 rocks along two of the MDS dimensions. These plots show that most of the MDS dimensions have clear psychological interpretations.

Dimension 1 corresponds to the lightness/darkness of the rocks' color, with dark rocks occupying the left side of Figure 4 and light rocks occupying the right. Dimension 2 corresponds to the average grain size of the rocks; rocks occupying the top of Figure 4 tend to have a large grain, while rocks at the bottom have little or no visible grain. Dimension 3 corresponds to the roughness/smoothness of the rocks' texture, with rough rocks occupying the right side of Figure 5 and smooth rocks occupying the left. Dimension 4 corresponds to the shininess of the rock, with dull rocks occupying the bottom of Figure *5* and shiny rocks occupying the top. Dimension 5 corresponds to the "organization" of the rocks; rocks occupying the left side of Figure *6* tend to be composed of fragments haphazardly glued together, whereas rocks to the right tend to have organized layers. Dimension 6 corresponds to the chromaticity of the rocks, with rocks occupying the top of Figure *6* having vivid, saturated colors and rocks occupying the bottom having dull, desaturated colors. Alternatively, dimension 6 may be interpreted in terms of warmness/coolness of color, with rocks occupying the top of the space tending to have warm colors and rocks at the bottom tending to have cool colors. For consistency, this dimension will simply be referred as "chromaticity", although there is likely truth to both interpretations.

The interpretations of dimensions 7 and 8 are not quite as clear-cut as the rest and were not actually rotated onto any hypothesized dimension ratings, leading Nosofsky et al. (2017) to label these "left-over" dimensions. Given the complexity of these rocks stimuli, it seems likely that people actually represent them using far more than 8 dimensions, so these left-over dimensions are probably amalgamations of several underlying psychological dimensions.

Meaningful patterns can still be identified, however. Notice that rocks occupying the left side of Figure 7 tend to have flat shapes, whereas rocks to the right tend to have more spherical or cubic shapes, indicating that this may be some sort of "shape" dimension. Also notice that rocks occupying the top of Figure 7 tend to be more green, whereas rocks at the bottom tend to be more red, indicating that this may be a "hue" dimension. Consistent with this interpretation, plotting dimension 6 (chromaticity) with dimension 8 as in Figure 8 produces a pattern reminiscent of the color circle. Starting from the top and moving clockwise, the color of the rocks shifts from red, to orange, to yellow, to green, to blue, to violet, and then finally back to red.

It has been shown in a number of experiments that these MDS dimensions can be used in conjunction with formal psychological models to predict human categorization behavior (Nosofsky, Sanders, & McDaniel, 2018; Nosofsky et al., submitted), indicating that they are good models of human psychological representations. The goal of the deep learning procedure will be to extract these representations automatically.

Data Splitting and Model Selection

The goal of training deep learning networks—or any other machine learning algorithm—is to find the model that minimizes some error function. In this case, the error function is the mean squared error (MSE) of the model's predicted MDS coordinates and the actual MDS coordinates. Deep networks have many free parameters whose values need to be set to minimize the error function. The deep learning literature makes a distinction between two types of free parameters. *Parameters* are free parameters that are optimized through a learning algorithm. Examples of parameters in deep networks are the connection weights between units, which are learned through backpropagation. *Hyperparameters* are free parameters that need to be set before

the learning process can begin, such as the number of layers or number of units in a deep network. Hyperparameters are typically optimized by conducting some sort of search through the hyperparameter space and finding the values that lead to the lowest error after training. Therefore, networks with different hyperparameter values need to be trained and compared to identify the network with the lowest error.

The naïve approach would be to train and compare networks using all 360 images from Nosofsky et al.'s (2017) set. However, deep networks may have millions of parameters, and thus nearly any network can trivially fit the training data, but it may overfit to noise and fail to generalize to new data. Because the goal is to create an automated system for extracting the psychological representations of novel stimuli, it is important to be able to estimate the networks' generalization performance and not just their training performance. To be able to do so, the data were split into three separate sets: the training set, the validation set, and the test set. Networks with different hyperparameter values were trained to minimize error on the training set, each network's performance on the validation set was evaluated, and the network with the lowest validation error was selected as the one with the best generalization error. However, because in this procedure the hyperparameter values are fit to the validation set, the network's validation error is likely lower than its true generalization error. Therefore, the network's error on the test set was then computed to gain an unbiased estimate of its ability to generalize to unseen data. Because none of the network's parameters or hyperparameters were fit to the test set, the network's output from the test images may be considered true predictions.

To summarize, the models' parameters were trained using the training set, the models' hyperparameters were fit to the validation set, and predictions to unseen data were generated using the test set. The training set was formed by randomly sampling 6 of the 12 rock tokens in

each category, and the remaining tokens were evenly split between the validation and test sets. Therefore, there were 180 images in the training set, and 90 in both the validation and test sets.

Deep Learning Models

I now describe several deep learning models that take images of rocks as input and predict the 8-dimensional MDS coordinates of those rocks as output, with the minimization of the mean squared error (MSE) between the model's output and the actual MDS values being used as the error function. All of the models in this section were implemented and trained using the Scikit-learn Python package (Pedregosa et al., 2011), the Keras Python package (Chollet & others, 2015), and the Tensorflow deep learning framework (Abadi et al., 2016). Hyperparameter optimization was conducted using the hyperopt Python package (Bergstra, Yamins, & Cox, 2013). Each of these models will be evaluated based on their generalization to the validation set (for a quick comparison between each model, refer to Table 2), and the model with the lowest validation error will be used to make predictions on the test set.

Null Model. The null model is a parameter-less model that always predicts the average values of the 8 MDS dimensions from the training set. The purpose of this model is simply to provide a baseline to compare the other models against. This model achieves a MSE of 6.67 on the validation set; any model that cannot perform better than this has not actually learned anything from the training data.

Scratch Model. The first real model I consider is a simple CNN trained "from scratch" on the rocks stimuli. This network takes 224x224 images as input, and consists of alternating convolutional and max pooling layers, followed by a series of fully-connected layers. For each of the fully-connected layers, ReLU, dropout, and batch normalization were used. The dropout rate was set to 0.5, and the batch normalization hyperparameters were left at their default values.

These layers fed into another fully-connected layer consisting of 8 linear units corresponding to the 8 MDS dimensions. The number of convolutional/pooling layers, the number of filters in each convolutional layer, the number of fully-connected hidden layers, and the number of units in each fully-connected hidden layer were hyperparameters fit to the validation set. The optimal values of these hyperparameters were found to be 4, 64, 1, and 128, respectively. *Adam* (Kingma & Ba, 2014) was used as the optimization algorithm, with its own hyperparameters left at their default values, except for the learning rate, which was found to have an optimal value of $10^{-2.82}$. To artificially increase the size of the training set, data augmentation was performed: training images were randomly flipped, rotated, cropped, and stretched/shrunk every time they were presented to the network. The network was trained for 200 epochs, but only the parameters from the epoch with the lowest validation error were saved. The training batch size was a hyperparameter fit to the validation set, and its optimal value was found to be 30.

This model achieved a MSE of 1.856 on the validation set. This is significantly better than the null model, but there is still room for improvement. The size of this network is quite small compared to most modern deep learning networks, but due to the small size of the training set, it would not be possible to train larger networks from scratch without overfitting. Therefore, the following models use larger networks trained on big data as a starting point. These networks have learned robust sets of visual features that are important for solving any computer vision task, such as edges, colors, and shapes. I explore different methods for adapting these features to my own task.

Feature Extraction Model. This model does not directly train CNNs, but rather uses them as a means for extracting features from images that can then be used as input for other machine learning algorithms. This is the approach previous researchers have taken in examining

the relationship between psychological and CNN representations, as outlined in the Related Work section. Moreover, this is a widely-used approach in the deep learning literature. It has been shown that the features learned by a CNN trained to perform one task using one data set can be used to solve new tasks using new data sets (Donahue et al., 2014; Sharif Razavian, Azizpour, Sullivan, & Carlsson, 2014). This approach is especially attractive in cases where limited training data is available because features can be extracted from CNNs trained using big data, and these features will generally be much more robust and complex than could be learned by a CNN trained solely on a smaller data set.

To implement this model, 3 well-known CNNs were downloaded from the internet: VGG16 (Simonyan & Zisserman, 2014), ResNet50 (He et al., 2016), and InceptionV3 (Szegedy et al., 2016). All of these networks were pre-trained to perform image classification on the ILSVRC data set. These networks were chosen not only because of their proven success at solving computer vision problems but also because they all represent different types of network architectures, and thus should yield different types of features. Features were extracted by providing the rock images as input to each of the networks, then extracting the activations from the final pooling layers of each network, which immediately preceded their final fully-connected layers and classification layers. Global average pooling was used to convert these 3-dimensional activations into 1-dimensional vectors for each image. These features were then used as input to a ridge regression model, which was trained to predict the MDS coordinates of the images in the training set.

This model had two hyperparameters: the pre-trained CNN used to extract the features and the regularization parameter used by the ridge regression model. It was found that ResNet50 and a regularization of 494 led to the lowest validation error. This model achieved a MSE of

2.132 on the validation set, which is significantly better than the null model, but not quite as good as the scratch model, indicating that additional training is necessary to align the representations used by these off-the-shelf networks with the MDS dimensions. The following model attempts to combine the benefits of task-specific training with pre-trained representations.

Transfer Learning Model. As discussed previously, CNNs trained on large data sets such as the ILSVRC will learn much more robust and complex features than CNNs trained on small data sets such as the 360-rocks set. However, CNNs do not necessarily need to be trained from scratch. CNNs trained on big data can be used as a starting point for training on a smaller data set, a technique known as *transfer learning* (Yosinski, Clune, Bengio, & Lipson, 2014). For this model, I used ResNet50, the network that yielded the best-performing features from the previous model, as a starting point for directly training a CNN to produce a rock's MDS coordinates.

I took the same approach as before of extracting the activations from the network's final pooling layer and applying global average pooling to produce feature vectors for each image, but this time the features were used as input to a series of fully-connected layers that were appended to the base CNN architecture. Similar to the scratch model, for each of these layers, ReLU, dropout, and batch normalization were used. These layers fed into another fully-connected layer consisting of 8 linear units corresponding to the 8 MDS dimensions. Adam was again used as the optimization algorithm, and data augmentation was performed during training. The model was trained until validation error stopped decreasing for at least 20 epochs, or for a maximum of 500 epochs. Only the newly-added fully-connected layers were trained in this model; the parameters in the base CNN architecture were left unchanged. This model had the following

hyperparameters: the number of hidden layers, the number of units in each hidden layer, the training batch size, and the learning rate, which were set to 2, 256, 90, and $10^{-2.22}$, respectively.

This model achieved a MSE of 1.494 on the validation set, which was not only significantly better than the null model, but also significantly better than the scratch model or the feature extraction model. It seems that combining task-specific training with pre-training on big data is better than doing either one alone.

Fine-tuned Model. The transfer learning model is able to take advantage of low-level visual features learned by a CNN trained on big data while learning its own higher-level features relevant to my particular task. It is possible that performance could be improved even more by also adapting the low-level features to my task using a procedure known as *fine tuning* (Yosinski et al., 2014). To implement this model, I took the transfer learning model from the previous section and trained it for an additional 500 epochs, saving only the parameter values at the epoch with the lowest validation error. This time all of the network's parameters were trained, including those in the base CNN architecture. Because the parameters in the early layers were expected to already be close to their optimal values, stochastic gradient descent with a low learning rate and high momentum (0.0001 and 0.9, respectively) was chosen as the optimization algorithm. This model had no additional hyperparameters compared to the previous one.

This model achieved a MSE of 1.330 on the validation set. It should be pointed out that it was not clear ahead of time that this model would outperform the transfer learning model on the validation set, because training all of ResNet 50's parameters on such a small training set could have easily led to overfitting. Indeed it is possible that even better generalization could have been achieved if additional hyperparameters had been introduced to control exactly which layers would be fine-tuned, but this will need to be explored in future work.

Ensemble Model. Networks with the same architectures and hyperparameters may converge to different minima in the error space if their parameters are initialized to different random values, and it has been shown that combining the outputs of multiple networks into an ensemble usually yields better results than using any individual network (Hansen & Salamon, 1990). To implement my own ensemble model, 10 separate instances of the transfer learning model were trained and fine-tuned, all using the same procedures and hyperparameters listed above. The outputs of these models were then averaged to produce the final predictions on the validation set.

This model achieved a MSE of 1.298 on the validation set, which was only slightly better than the fine-tuned model. This result could indicate that training a full ensemble of CNNs is simply not worth the time or the effort for the present task. It could also be, though, that the predictions of the individual networks were too similar to each other to gain any benefits from averaging. A greater variety of predictions could be achieved by having each network use one of the top 10 best-performing sets of hyperparameters, rather than each network using the same single best-performing set. Future research will need to explore this idea and other more sophisticated ensembling schemes (e.g., S. Lee, Purushwalkam, Cogswell, Crandall, & Batra, 2015).

Generalization to Test Set

The ensemble of CNNs model described above was identified as the model with the best validation error, achieving a MSE of 1.298 and an R^2 of 0.780. While promising, this is likely an overestimate of true generalization performance because the ensemble was fit to the validation data. To get an unbiased estimate of the model's generalization ability, it needs to be evaluated on the test set. To reiterate, none of the networks' parameters or hyperparameters were

manipulated to decrease error on the test set, so predictions on the test set are true predictions of unseen data. The ensemble achieved a MSE of 1.355 and an R^2 of 0.767 on the test set. The fact that the ensemble accounts for over 75% of the variance in both the validation and test sets provides converging evidence that CNNs can be trained to extract psychological representations from novel stimuli. Code for training this model can be found on this dissertation's associated website (https://osf.io/d6b9y/), and a 3-network version of this ensemble can be also accessed, allowing the user to upload images of rocks and download their MDS representations as a .csv file.

Predictions of MDS Dimensions. Figure 9 - Figure 16 display scatterplots of the actual MDS values of the rocks from the test set against the values predicted by the ensemble of CNNs. Recall that there were 90 rocks in the test set, so there are 90 data points in each figure. As can be seen, the correlation between the ensemble's predictions and the actual MDS values is very high for most of the dimensions. The CNNs perform the best on the lightness and chromaticity dimensions, which is unsurprising given that these dimensions reflect low-level color information, and it is probably the case that even simple machine learning models could extract these dimensions. It is also probably unsurprising that the CNNs perform less well on the "shape" dimension given that this dimension does not have a clear interpretation, and it is likely an amalgamation of several underlying psychological dimensions. Indeed, the fact that the networks are able to make even somewhat accurate predictions for this dimension is remarkable and indicates that it does hold some meaning, even if that meaning is not immediately apparent to human observers.

What may be surprising about the ensemble's predictions is that the CNNs perform almost as poorly on the roughness dimensions as the shape dimension, even though the former

seems to have a clearer interpretation. Inspection of the rocks that the CNNs mis-predict reveals that there are several rocks located on the left side of the MDS space that actually have textures that seem rougher than their MDS coordinates would suggest. In particular, consider Anthracite 3, the shiny gray rock located at approximately (0, 4) in Figure 11, and Dolomite 5, the beige rock located at approximately (-4, -1) in Figure 11. According to the MDS solution, Anthracite 3 should have a texture that is neither particularly rough nor smooth, but the jagged surfaces and sharp corners suggest a rough texture more in line with the ensemble's prediction. Similarly, according to the MDS solution, Dolomite 5 should have a very smooth texture, but while the wavy surface of this rock does not seem especially rough, it does not seem especially smooth either, which is more in accordance with the ensemble's prediction. All of this is to say that some of the mis-predictions for this dimension seem not to be due to failures in the CNNs, but rather due to failures in the MDS solution itself. Given that it was derived from an incomplete similarity matrix, it is only to be expected that the MDS solution would have noise. Indeed, the limitations of this MDS solution will be a theme throughout the rest of this dissertation.

Predictions of Similarity Judgments. While the previous analysis examined whether the CNNs can predict the values of individual MDS dimensions, another important question is how the MDS and CNN representations compare on their ability to predict human similarity judgments. Even if the networks fail on certain dimensions, as long as the CNN representations properly place similar items close together in space and dissimilar items far apart, they may be useful for modeling categorization behavior. Because of the sparse nature of Nosofsky et al.'s (2017) similarity data, I choose to examine average similarity judgments between subtypes of rocks, rather than the similarities between pairs of individual rocks. Specifically, all of the similarity judgments from Nosofsky et al.'s (2017) data set that involved pairs of rocks from

only the test set were used to compute the average similarity between each pair of rock subtypes. The average distances between each pair of subtypes was then computed using both the MDS and CNN representations, again only using the rocks from the test set, and equation (1) was used to derive predicted similarities.

Figure 17 displays scatterplots of these observed and predicted similarities, with the MDS-predictions in the top panel, and the CNN-predictions in the bottom. Within each plot, closed circles represent the average similarity of all pairs of tokens within a particular subtype, open circles represent the average similarity of all pairs of tokens between two particular subtypes belonging to the same high-level category (igneous, metamorphic, or sedimentary), and crosses represent the average similarity of all pairs of tokens between two particular subtypes belonging to different high-level categories. Overall, the MDS representations are able to explain 65.8% of the variance in this data[4], and the CNN representations are able to explain 55.1%. Given that the MDS model was fitted directly to the similarity judgments, it is not surprising that it provides a quantitatively better fit to the data. However, inspection of the scatterplots indicates that the ensemble of CNNs makes predictions that are qualitatively very similar to those of the MDS model, indicating that the CNNs have learned to generalize the MDS representations. Although neither model fits the data very well due to the high amount of noise in the similarity judgments, one virtue of both models is that they are able to correctly predict that many subtypes of rocks, even ones from different high-level categories, are more similar to each other than some subtypes are similar to themselves. This further reaffirms the intuition that rocks have a dispersed category structure that violates the family-resemblance principle.

[4] When this analysis is conducted using the full set of 360 rocks, the MDS solution is able to account for 96.6% of the variance in the data. Given the smaller sample size and the high amount of noise in the similarity judgments for individual pairs of rocks, it is not surprising that the MDS solution is able to explain less of the variance in the present analysis.

Even where the models fail, they tend to fail in similar ways. For example, both models underestimate the very high within-subtype similarity of several subtypes such as obsidian and marble. This may be because neither set of representations capture certain idiosyncratic dimensions that may increase within-subtype similarity. Obsidian, for example, can be recognized by its shiny scalloped surfaces, and while marble is generally white, many examples have dark veins running through them. It may be necessary to explicitly add such dimensions to both the MDS and CNN representations to allow them to better capture psychological representations. Alternatively, it may be necessary to build prior knowledge into the models. Both obsidian and marble are often referenced in popular culture, so some participants probably recognized when they were presented with two examples belonging to one of these subtypes and reported higher similarities.

The MDS and CNN representations also underestimate the between-subtype similarity of granite and diorite (this data point is represented by the open circle at approximately (7.1, 5.7) in both panels). This is another case where the models may fail because they lack certain idiosyncratic dimensions; both granite and diorite tend to have dark freckles with a lighter background. Both models also underestimate the similarity between obsidian and sandstone (this data point is represented by the cross at approximately (4.6, 2.6) in both panels). This is a case where there is probably noise in the data, and the model predictions seem more sensible; obsidian tends to be black and shiny, while sandstone tends to be chromatic and dull. Another interesting case is that both models overestimate the similarity between pumice and sandstone (this data point is represented by the cross located at approximately (1.9, 3.1) in both panels). The models likely predicted some similarity between these subtypes due to similar lightness and chromaticity values, but people likely reported little similarity because they focused more on

certain dimensions that the models did not have access to. Namely, all of the test examples of pumice had holes, and two of the three examples of sandstone had stripes, and these unique features likely made these subtypes seem very dissimilar to people. Again, explicitly adding such idiosyncratic dimensions to the MDS and CNN representations should allow them to better capture human judgments.

Discussion

Overall, the ensemble of CNNs does an excellent job at predicting the MDS coordinates of rocks it was not trained on, and the MDS and CNN representations make comparable predictions regarding the similarities between subtypes of rocks, indicating that they should also make comparable predictions regarding categorization behavior. Furthermore, many of the CNNs' mis-predictions seem to be due to issues with the data rather than issues with the networks themselves. These facts lend promise to the idea that CNNs may be used to automatically extract psychological representations from images.

While I have emphasized that it is important to test the models on untrained stimuli to ensure that the models are generalizing to novel input and have not overfitted to the training data, there is a sense in which the test set I have used here is not completely independent from the training or validation sets since all the sets came from the same MDS solution. It is not clear that the same dimensions would emerge if the MDS analyses were conducted again using a new set of rocks, even if those rocks were sampled from the same subtypes used in the original set. If different MDS dimensions did emerge for different sets of rocks, then the CNNs would not actually be able to generalize to new stimuli, in spite of the results reported here. Therefore in the next section, I describe an MDS analysis using a new set of rocks to see whether the same

dimensions would emerge as in Nosofsky et al.'s (2017) data set and to assess whether the CNNs are able to generalize to these completely novel stimuli.

Experiment 1

This experiment is a conceptual replication of Nosofsky et al.'s (2017) study, using a new set of rock images sampled from the same 30 subtypes of rocks. Again, these subtypes are listed in Table 1, and they are meant to be reprehensive of the types of rocks found in introductory geology textbooks. The goal was to collect similarity ratings between pairs of rocks and direct ratings along hypothesized dimensions for each individual rock, so that MDS analyses could be conducted. As in Nosofsky et al.'s (2017) study, the dimensions that were hypothesized to emerge in the MDS analysis were (1), lightness/darkness of color, (2) average grain size, (3) smoothness/roughness, (4) shininess, (5) organization, and (6) chromaticity. The purpose of this experiment was to see if these dimensions would emerge again if MDS analyses were conducted on a new set of rocks and to see if the ensemble of CNNs described in the previous section could predict the MDS dimensions or the similarity relations of these novel stimuli. The stimuli, collected data, and derived MDS solution from this experiment can be found on this dissertation's associated website (https://osf.io/w64fv/).

Method

Participants. The participants were 125 students from introductory psychology courses at Indiana University, Bloomington. Participants received credit towards a course requirement. All participants reported normal or corrected-to-normal vision and no expertise in geology. Of these participants, 85 provided similarity judgments, 20 provided direct ratings for the first three hypothesized dimensions (lightness/darkness of color, average grain size, and smoothness/roughness), and 20 provide direct ratings for the second three hypothesized dimensions (shininess, organization, and chromaticity).

Stimuli. The stimuli were 120 images of rocks belonging to the same 30 subtypes used by Nosofsky et al. (2017), although none of the individual images were repeated. There were 4 individual tokens in each subtype. Some of these new images were obtained through web searches, while others were taken from Meagher et al. (in press). Photoshopping procedures were used to remove backgrounds and idiosyncratic markings such as text labels from the images.

The stimuli were presented on a 23-in. LCD computer screen with a white background. Each rock picture was approximately 2.1 in. wide and 1.7 in. tall, and participants sat approximately 20 in. from the computer screen, so each rock picture subtended a visual angle of approximately 6.0° x 4.9°. Each image was constrained to have a horizontal resolution of 800 pixels, but the height was allowed to vary to preserve aspect ratio. All stimuli and instructions were displayed on a computer screen using MATLAB and Psychtoolbox (Brainard, 1997).

Similarity-Judgments Procedure. Participants were shown pairs of rock pictures and were instructed to judge the similarity of the rocks on a scale from 1 (most dissimilar) to 9 (most similar). On each trial, two subtypes were randomly selected (both subtypes could be the same), and then one token was randomly selected as a representative within each subtype (the same token could not be selected twice when the subtypes were the same). One token was placed on the left side of the screen, and the other was placed on the right. The participant gave their similarity judgment for the pair using the computer keyboard. This procedure was repeated for all 435 unique pairs of the 30 rock subtypes, as well as all 30 within-subtype comparisons, for a total of 465 trials. Participants were not instructed on what features to use to make their judgments, but they completed 5 practice trials to get a sense of the types of stimuli they would see.

Dimension-Ratings Procedure. Participants were shown one of the 120 rocks on each trial and were asked to provide a rating on a 1-9 scale along one of the hypothesized dimensions in Table 3. Responses were entered using the computer keyboard. Participants provided ratings for all 120 rocks for one dimension before moving on to the next dimension. The presentation order of the dimensions and of the individual rocks was randomized for each participant. To promote a consistent scale across participants for each dimension, the 1-9 scale was shown at the bottom of the screen on each trial, with labeled anchor pictures at the middle and extreme ends of the scale. For example, in the "lightness of color dimension" a picture of a dark black rock labeled "Darkest" was displayed between 1 and 2 on the scale, a picture of a gray rock labeled "Medium" was displayed at 5, and a picture of a bright white rock labeled "Lightest" was displayed between 8 and 9 on the scale. The text labels used for each dimension can be found in Table 3.

Results

Before conducting any MDS analyses, the data from individual participants were compared to group-averaged data to identify outliers. For each participant that provided similarity judgments, a 30x30 similarity matrix was constructed, where cell *i, j* indicated the participant's similarity between rock subtypes *i* and *j*. Correlations were then computed between the individual participant similarity matrices and the group similarity matrix averaged across all participants, and six participants were identified as having especially low correlations ($r < .4$), so the average similarity matrix was recomputed with the data from these participants removed. Similarly, for both sets of direct dimension ratings, correlations were computed between the group-average ratings and the ratings of each individual participant. For the first set of dimensions, one participant was identified as having a lower correlation than the others ($r = .63$),

and for the second set of dimensions, three participants were identified as having lower correlations ($r < .6$), so the average dimension ratings were also recomputed with these participants' data removed. The averaged similarity matrix was then used to fit an 8-dimensional MDS solution, using equations (1) and (2) from above. The MDS space was then rotated onto the averaged dimension ratings using equation (3), in the same manner as described previously.

Predictions of MDS Dimensions. Figure 18-Figure *21* display the rotated MDS dimensions, and Figure 23-Figure *30* display scatterplots between these MDS dimensions and the 8 predicted dimensions from the ensemble of CNNs. None of the CNNs parameters or hyperparameters were fit to these data in any way, so these represent true predictions from the ensemble. Inspection of these figures reveals that as in Nosofsky et al.'s (2017) 360-rock MDS solution, dimensions 1, 2, 4, and 6 are interpretable in terms of lightness/darkness, average grain size, shininess, and chromaticity, respectively. These interpretations are corroborated by the high correlations between these MDS dimensions and the direct dimension ratings, as reported in Table 3. Furthermore, the correlation between these MDS dimensions and the dimensions predicted by the ensemble of CNNs are high[5] (see Figures Figure *23*, Figure *24*, Figure *26*, and Figure *28*), indicating that the ensemble is able to generalize these dimensions to rocks that were not even included in the MDS solution the networks were trained on. There are a few noteworthy discrepancies between the CNN-predicted values and the MDS-derived values, however. Notice, for example, that the CNNs underestimate the grain size and overestimate the shininess of Conglomerate 1, which is the blue rock with large embedded pebbles at approximately (4, 0) in

5 The slopes of the CNN predictions are higher than the slopes of the MDS predictions, reflecting the different scales of the 360-rock and 120-rock MDS solutions (compare the axes of Figure 4Figure *7* with those of Figure 18Figure *21*).

Figure 24 and (0, 4) in Figure 26; it may be the case that the CNNs are misinterpreting the embedded white pebbles in this rock as shiny spots.

The interpretation of dimensions 3 and 5 for the 120-rock MDS solution are not quite as clear-cut as they were in the 360-rock MDS solution (see Figures Figure *25* and Figure *27* and Table 3). While it does seem to be generally true that rocks on the right side of Figure 19 are rougher than those on the left side, there are many exceptions, and the correlation between this MDS dimension and the direct roughness ratings was modest. Similar observations can be made for disorganized versus organized rocks in Figure 20. Given that these dimensions failed to strongly replicate from the 360 rock MDS solution, it is unsurprising that correlation between them and the CNN-predicted dimensions was low (see Figures Figure *25* and Figure *27*), casting some doubt on the CNNs' ability to generalize to new sets of rocks.

It should be pointed out, though, that even in Nosofsky et al.'s original study, the correlation between the direct smoothness/roughness and organization ratings and their associated MDS dimensions were only .654 and .694, respectively, indicating that these may not have been the best interpretations for those dimensions in the first place. While it was discussed previously that there may be noise in these dimensions due to insufficient similarity judgment data, it is also probably the case that 8 MDS dimensions is an insufficient number of dimensions for truly capturing psychological representations. Dimension 3 in the 360 rock MDS solution likely reflects several underlying texture-related dimensions including not only roughness/smoothness, but also features such as whether the surface of the rock is flat, wavy, or bumpy. Similarly, Dimension 5 in the 360-rock MDS solution likely reflects several underlying organization-related dimensions including not only whether the rock is made of fragments or layers/stripes, but also whether the fragments are angular or rounded, and whether the

layers/stripes are straight or curved. These distinctions are actually important for distinguishing certain categories of rocks. Breccia and conglomerate, for example, are both composed of disorganized fragments, but the fragments in breccia are angular, and the fragments in conglomerate are round. It may be that there were not enough rocks in the 120-rock data set for these subtle features to factor into participants' similarity judgments, causing a failure to strongly replicate the same dimensions from the 360-rock data set.

Given that dimensions 7 and 8 in the 360-rock MDS solution did not have clear interpretations, it is surprising to see that similar dimensions emerged again in this 120-rock MDS solution (compare Figure 7 with Figure 21). Notice that the rocks on the left side of Figure 21 tend to be flat, while the rocks on the right side tend to be more spherical or cubical, indicating that shape again influenced participants' similarity ratings. And notice that there are many blue, purple, and red rocks at the bottom of Figure 21, while there are more yellow, brown, and green rocks at the top, indicating that hue was also a factor. In support of this interpretation, plotting dimensions 6 and 8 together (Figure 22) once again forms a loose color circle: starting from the top left and moving clockwise, the colors shift from red, to orange, to yellow, to green, to blue, to purple, and back to red (with the achromatic rocks lying to the lower-left of the plot). And while the correlations between these MDS dimensions and the CNN-predicted dimensions are relatively modest (Figures Figure *29* and Figure *30*), the fact that the networks were able to generalize at all along these nebulous dimensions is quite impressive. Moreover, the fact that these dimensions emerged in both the 360-rock and 120-rock MDS solutions indicates that they are psychologically meaningful and are not just "leftovers." Future research will need to find solid interpretations of these dimensions, perhaps by collecting direct ratings of shape and hue-related dimensions and rotating the space again.

Predictions of Similarity Judgments. Figure 31 displays scatterplots of the observed and predicted similarity judgments between pairs of subtypes of rocks, with the MDS predictions in the top panel, and the CNN predictions in the bottom panel. The data in these scatterplots were derived using the same methods as those used for Figure 17. Again, the MDS model was fitted directly to these similarity-judgment data, whereas the CNNs were not fit to these data in any fashion. The MDS representations are able to explain 96.8% of the variance in the similarity judgments, while the CNN representations are able to explain 76.6% of the variance. The higher R^2 values found in this analysis compared to the analysis displayed in Figure 17 reflect the higher sample sizes in each subtype and decreased noise for individual similarity judgments. Naturally, the fit of the zero-parameter CNN model is worse than that of the high-parameter MDS model. The important point is that, without estimating any free parameters, the CNNs can explain a reasonably high proportion of the variance in the similarity judgment data. This success is achieved despite the fact that not all of the dimensions from the 360-rock MDS solution strongly replicated in the 120-rock MDS solution. The reasonably good fit of the CNN model suggests that the CNNs representations correctly place similar items close together and dissimilar items far apart.

As can be seen, the MDS solution somewhat underestimates the high within-subtype similarity of a few subtypes including pumice, gabbro, and rock gypsum, and this pattern is even more exaggerated in the CNN predictions. Again, this is likely because neither set of representations capture some of the features or sets of features unique to these subtypes: all the examples of pumice had holes, all of the examples of gabbro had a similar dark brown color with light brown freckles, and all the examples of rock gypsum had thin crystals layered on top of each other. Some of the CNNs' other mis-predictions are illuminating for diagnosing problems

with the networks. For example, the network underestimates the similarity between basalt and pumice, likely because the networks did not recognize that, like pumice, some examples of basalt had holes. This again indicates that it may be beneficial to explicitly train the CNNS to detect holes and other idiosyncratic features. The networks also overestimate the similarity between schist and conglomerate. Schist can be recognized by its shiny surfaces, so this may again indicate that the networks mistook the white pebbles embedded in Conglomerate 1 as shiny spots. It may be necessary to train the networks on more examples similar to Conglomerate 1 so that they can better learn to differentiate shiny surfaces from simply white surfaces.

Discussion

The MDS solution for the 120-rock data set again yielded dimensions that were clearly interpretable in terms of lightness/darkness of color, average grain size, shininess, and chromaticity, indicating that these are important psychological dimensions across a wide variety of sets of rocks. The smoothness/roughness and organization dimensions did not emerge as clearly in the previous 360-rock analysis, perhaps indicating that these dimensions actually reflect several different underlying psychological dimensions that may manifest differently in different MDS analyses. The fact that similar shape- and hue-related dimensions emerged in both the 360- and 120-rock data sets indicates that these actually are important psychological dimensions, and future research should try to find concrete interpretations for them.

The fact that the CNNs are able to account for over 75% of the variance in the similarity judgments for the 120 rocks lends promise to the idea that they could be used in conjunction with cognitive models to predict categorization behavior. This cannot be known for certain, of course, until the idea is actually tested, so in the following section I describe a categorization experiment

using these 120 rocks and compare the MDS and CNN representations on their ability to predict human categorization behavior.

Experiment 2

This categorization experiment was conducted to compare the CNN and MDS representations on their ability to predict human categorization behavior when used in conjunction with a formal psychological model. In particular, I used the Generalized Context Model (Nosofsky, 1986, 2011) because previous work has shown that it can provide good first-order predictions of human rock categorization (Nosofsky, Sanders, & McDaniel, 2018; Nosofsky et al., submitted).

There were three conditions in this experiment. Two of these conditions were conceptual replications of experiments conducted by Nosofsky, Sanders, & McDaniel (2017): the igneous condition, in which participants were tasked with learning the 10 subtypes of igneous rocks, and the mixed condition, in which participants were tasked with learning a mixture of igneous, metamorphic, and sedimentary rocks (see Table 4 for the specific subtypes used in the mixed condition). The third condition was the metamorphic condition, in which participants learned the 10 subtypes of metamorphic rocks. This design allowed me to test whether the representations could account for performance differences for the same subtypes between different conditions, based on changes in between-category similarity relations across the conditions. Specifically, because diorite has high similarity to granite in the igneous condition, but has no close neighbors in the mixed condition, I hypothesized that diorite would be categorized correctly more often in the mixed condition than in the igneous condition. Similarly, I hypothesized that obsidian and anthracite, which are both dark, shiny rocks, would be categorized correctly more often when they were separated into the igneous and metamorphic conditions, than when they were together in the mixed condition.

Method

Participants. The participants were 133 members of the Indiana University Bloomington community. Participants were compensated $10 with a possible $2 bonus for scoring at least 60% correct during the test phase of the experiment. There were 8 participants who did not achieve this criterion, and their data were excluded from further analyses, leaving 41 participants in the igneous and mixed conditions, and 43 in the metamorphic condition.

Stimuli. The stimuli were the same 120 rocks used in Experiment 1. Within each subtype of rock, the first two tokens were selected as training stimuli, and the second two tokens were selected as transfer stimuli. Because there were 10 subtypes in each condition, there was a total of 20 training stimuli in each condition and 20 novel items presented at time of test.

Procedure. Each participant was randomly assigned to one of the 3 conditions: igneous, metamorphic, or mixed. The experiment was divided into a training phase and a test phase. The training phase consisted of 6 blocks of trials. On each trial, participants were asked to categorize a single training item using the keyboard, and they were given feedback after entering their answer. The feedback always told participants the correct answer (e.g., "Correct, Andesite!" or "Incorrect, Basalt!"). Each of the 20 training items was presented twice every block in random order. The test phase consisted of 4 blocks of trials. In this phase, each training and transfer item was presented once every block in random order, and no feedback was given for the transfer items. To keep participants engaged in the task, feedback was given for each training item once in the first two test blocks and once in the second two test blocks.

Formal Model

The Generalized Context Model (GCM) was fit to the categorization data from the test phase of the experiment[6]. This model assumes that people store exemplars of categories in memory and that stimuli are categorized according to how similar they are to these exemplars. I start by describing a full version of the model applied to the present experiment, although many of the ensuing analyses will consider constrained versions in which special cases of the model are applied.

Formally, the GCM states that the probability that item *i* is categorized into category *J* is found by summing the similarity of *i* to all training exemplars of category *J* and then dividing by the summed similarity of *i* to all exemplars of all categories:

$$P(J|i) = \frac{b_J\left(\sum_{j\in J} s_{ij}\right)^{\gamma}}{\sum_K b_K\left(\sum_{k\in K} s_{ik}\right)^{\gamma}} \quad (4)$$

where, s_{ij} denotes the similarity of item i to exemplar j; b_J $\left(0 < b_J < 1,\ \sum b_J = 1\right)$ is the response bias for category J; and γ is a response-scaling parameter. The response bias parameters allow for asymmetric category confusions. For example, if a rock is equally similar to the training exemplars of granite and diorite, but the response bias is high for granite and low for diorite, then it is more likely that the rock will be categorized as granite. The default setting of the response-scaling parameter is γ=1; in this case, the observer responds by probability-matching to the relative summed similarities of the categories. As γ grows greater than 1, the observer responds more deterministically with the category that yields the largest summed similarity.

[6] Modeling the time course of learning in the training phase will be an important avenue for future research.

The similarity between item *i* and exemplar *j*, s_{ij}, is given by

$$s_{ij} = \begin{cases} e^{-c_b d_{ij}}, \text{if } i \text{ and } j \text{ belong to different categories} \\ e^{-c_w d_{ij}}, \text{if } i \text{ and } j \text{ belong to the same category} \end{cases} \quad (5)$$

where d_{ij} is the psychological distance between item i and exemplar j, and c_b and c_w are sensitivity parameters that determine the rate at which similarity declines with distance. While previous applications of GCM have assumed a single sensitivity parameter, here different similarity gradients are allowed for between- and within-category comparisons. The reason, as discussed previously, is that many categories of rocks have distinctive features that are not part of the MDS or CNN representations but may nonetheless cause increased within-category similarity, such as pumice's holes or marble's dark veins. When $c_b > c_w$, the model is able to capture some of this increased within-category similarity even though it does not have access to these distinctive features. Further justification for this parameterization of the model is given in (Nosofsky et al., submitted), but it should be emphasized that future research will need to expand the psychological space with more dimensions to construct a more complete theory of rock categorization.

Let x_{im} and x_{jm} denote the values of item *i* and exemplar *j* on dimension *m* in an *M*-dimensional psychological space. The psychological distance between item *i* and exemplar *j*, d_{ij}, is computed as the weighted Euclidean distance:

$$d_{ij} = \sqrt{\sum_{m=1}^{M} w_m \left(x_{im} - x_{jm}\right)^2} \quad (6)$$

where w_m ($0 < w_m < 1,\ \sum w_m = 1$) is the attention weight given to dimension *m*. These attention weights stretch and shrink psychological space to amplify differences along dimensions relevant for categorization or to reduce difference along irrelevant dimensions. The attention

weights have been crucial for fitting GCM to prior experiments using artificial category structures in which some dimensions may be highly relevant and others irrelevant; however, in the present naturalistic domain, all the dimensions tend to be relevant for categorization, so the attention weights may play a less dramatic role.

Results

Fitting averaged category data. The bars in Figure 32 display the mean proportion of correct categorization decisions during the test phase as a function of condition and item type (old training or new transfer items). Inspection of this figure reveals that participants in all 3 conditions correctly categorized the training items nearly 100% of the time, indicating that errors on the transfer items were due to their similarity to members of competing categories and not simply a failure to learn the training items. The figure also indicates that the categories in the mixed condition were somewhat easier to learn than those in the other conditions—transfer items in the mixed condition were correctly categorized nearly 80% of the time, while they were correctly categorized only about 60% of the time in the other conditions (chance performance is 10%).

Figure 33 presents a more fine-grained view of the data. The bars displays the mean proportion of correct categorization decisions for the transfer items as a function of condition and the individual categories of rocks (performance for the training items was near ceiling for every category). Inspection of this figure reveals that within each condition, the categories varied in difficulty, and many of these patterns are similar to those observed in prior work (Meagher et al., submitted; Nosofsky, Sanders, & McDaniel, 2018; Nosofsky et al., submitted). For example, generalization performance was quite high for obsidian and pumice, while it was low for rhyolite.

Figure 33 also reveals performance differences for the same categories in different conditions. As hypothesized, performance for diorite was higher in the mixed condition than in the igneous, and performance for anthracite was higher in the metamorphic condition than in the mixed. To test whether these differences were statistically significant, I performed two Bayesian independent samples t-tests using JASP (JASP Team, 2018), leaving all default settings intact. These tests compute Bayes factors (Kass & Raftery, 1995), which give the ratio of the likelihood of the data under the alternative hypothesis (nonzero difference between conditions) and under the null hypothesis (zero difference between conditions). The computed Bayes factors for diorite and anthracite were $4.144e \times 10^9$ and 1765, respectively, providing overwhelming evidence that there was a difference between conditions. However, while I also hypothesized that performance for obsidian would be higher in the igneous condition than in the mixed condition, there was actually very little difference between conditions in the data. A Bayesian independent samples t-test yielded a Bayes factor of 1.174, which provides very little evidence for a real difference between conditions.

To model these data I used a "baseline" version of the GCM that only allowed the sensitivity parameters, c_b and c_w, to vary. The attention weights were constrained to be equal to each other, as were the response biases, and the response-scaling parameter γ was set to 1. Viewing the results from this version of the model is useful because it gives a sense of how well the CNN and MDS representations perform at predicting human rock-classification *a priori*, without manipulating the space through the attention weights or response biases. I fitted this version of the model by calculating the proportion of correct categorization decisions in the test phase for all training items and all transfer items in each condition and category of rock, averaged across all participants, and then searching for parameter values that minimized the

MSE between the empirical observations and the model's predictions. I fitted two versions of this baseline model: one that used the CNN representations of the rocks as its input, and one that used the MDS representations of the rocks. The best-fitting parameters for the model using the CNN representations were $c_b = 0.947$ and $c_w = 0.665$, and for the model using the MDS representations they were $c_b = 1.205$ and $c_w = 0.855$.

In Figure 32 and Figure 33, GCM predictions using the CNN representations are indicated by circles, and GCM predictions using the MDS representations are indicated by crosses. Both models provide an excellent fit to the data collapsed across categories of rocks; the GCM+CNN model achieved an MSE of 0.0005 and $R^2 = 0.98$, while the GCM+MDS model achieved an MSE of 0.0002, $R^2 = 0.99$. At this coarse level of analysis, the models make nearly identical predictions. For the data divided into the individual rock categories the GCM+MDS model provides an overall better fit (MSE = 0.007, $R^2 = 0.82$) than the GCM+CNN model (MSE = 0.010, $R^2 = 0.75$), but for most categories their predictions were similar. Generally speaking, both models were able to predict which categories would be easy or hard. In addition, both models successfully predicted that performance for diorite would be higher in the mixed condition than in the igneous condition, and that performance for anthracite would be higher in the metamorphic condition than in the mixed condition.

There are cases in which the models fail, however. Both models tend to underestimate performance for pumice and slate, for example. This is likely because these categories had distinctive features recognized by the human participants but not captured in the CNN or MDS dimensions: all the examples of pumice contained holes, and all of the examples of slate were composed of thin gray sheets. While using separate sensitivity parameters for between- and within-category comparisons was intended to improve such shortcomings in the CNN and MDS

representations, it may be necessary to expand the representations with such idiosyncratic features to fully capture human behavior. Alternatively, it may be necessary to incorporate prior knowledge into the models; some participants may have already been familiar with pumice because it is often used as an exfoliant, or with slate because it is often used as a construction material. Another notable shortcoming of the models is that, in agreement with my initial hypothesis, they predicted performance for obsidian would be lower in the mixed condition than in the igneous condition (because it would be confused for anthracite in the mixed condition); however, this pattern was not seen in the observed data. Again, this may be because the representations were not able to capture certain distinctive features associated with obsidian, such the presence of scalloped surfaces, or because some participants had prior knowledge of that rock category (obsidian is often referenced in popular culture).

There are cases, though, where the predictions from the MDS and CNN representations differ dramatically, such as for pegmatite and rhyolite in the igneous condition, and for amphibolite in the metamorphic condition. It is difficult to say why the predictions mismatch without determining exactly which categories the models are confusing with each other, or what the predictions are for individual rock tokens. Therefore, in the analyses reported in the next section, I consider more sophisticated versions of GCM that are fitted to the entire confusion matrices in each condition.

Fitting individual rock data. The complete matrices of classification confusion data observed in each condition are reported in Figure 34-Figure 36. Each individual row in each matrix corresponds to an individual rock token, and each column corresponds to an individual category. The entry in row i and column J gives the conditional probability with which rock-token i was classified into rock-category J.

I fitted four versions of the GCM to these data, each using a different number of free parameters. Each model was fitted twice, once using the CNN representations as input, and once using the MDS representations. In the "baseline" version of the model, only the sensitivity parameters, c_b and c_w were allowed to vary, with the all the attention weights and all the category biases being constrained to be equal.[7] In the "baseline + weights" model, the attention weights were also allowed to vary, and in the "baseline + biases" model, the response biases were also allowed to vary. In the "full" version of the model, the sensitivity, attention weight, and response bias parameters were all allowed to vary. Unlike the sensitivity parameters, the attention weights and response biases were allowed to vary across conditions (because the categories varied across the conditions). There were 3 conditions, 8 psychological dimensions, and 10 categories in each condition, and the sum of the attention weights and the sum of the response biases can both be constrained to equal 1. This means that there were $3 \times (8-1) = 21$ free attention weight parameters, and $3 \times (10-1) = 27$ free category response bias parameters. For each of these models, an equal-sensitivity version such that $c_w = c_b$ was also fitted to the data.

All models were fitted to the data using a maximum-likelihood criterion. Because the models vary in their number of free parameters, the fits were compared using the Bayesian Information Criterion (BIC; Schwarz, 1978), which penalizes models with more free parameters. The BIC fit of a model is given by

$$\mathrm{BIC} = -2\ln L + P\ln(N) \qquad (7)$$

[7] Because the response-scaling parameter γ is mainly important when fitting the data of individual participants, and here I am fitting group-averaged data, γ was constrained to be 1 in all models. Fitting individual participant data will be an important topic for future research.

where L is the maximum likelihood of the data given the model, P is the number of free parameters, and N is the sample size. The model with the smallest BIC is considered to provide the most parsimonious account of the data.

Table 5 presents the negative log-likelihood and BIC fits of each GCM model, using both the CNN and MDS representations (lower negative log-likelihood and BIC scores indicate better fits). This table reveals a number of important points. First, even with its large number of free parameters, the full GCM is still considered the most parsimonious account of the data using both representations. This result indicates that the response bias and attention weight parameters may in fact be important for explaining the categorization of these rocks stimuli; however, given the preliminary nature of the psychological feature space, the role of these parameters should be interpreted with caution (the best-fitting parameters of this model can be found in Table 6). A second important point is that constraining $c_w = c_b$ dramatically worsens the model fit: even the full version of the GCM with equal sensitivities (with 49 free parameters) yields a worse negative log-likelihood fit than does the baseline model that allows unequal sensitivities (a 2-parameter model). This reinforces the point that both the CNN and MDS representations are missing some key dimensions to which humans may attend in classifying the stimuli. Finally, for each model, the MDS representation yields a better BIC fit than the CNN representation, indicating that the CNN representations are still not quite capturing similarity relations among the rocks as well as the MDS representations.

Figure 34, Figure 35, and Figure 36 show, along with the observed individual rock-token classification data, the predictions from the full GCM model using both CNN and MDS representations. Darker cell shadings indicate higher proportions. By looking at these shadings, it can be seen at a glance where the models agree or disagree with the empirical data and with each

other. For instance, it can be seen that both models correctly predicted high accuracy for obsidian in the igneous condition, but that in the mixed condition the models (especially the CNN model) were more likely than humans to categorize obsidian as anthracite. Inspection of these matrices reveals cases where fits to the averaged data may be misleading. For instance, while Figure Figure 33 indicates that the MDS model fit to the averaged data accurately predicts performance for Amphibolite in the Metamorphic condition, Figure 35 reveals that the MDS model fit to the entire confusion matrix actually greatly overestimates performance for Amphibolite 3 and underestimates performance for Amphibolite 4, and its predictions are more similar to the CNN model than the averaged results may suggest (a similar pattern of results can be seen for gneiss).

Inspection of mis-predictions involving the individual rocks is informative for understanding where the CNN and MDS representations fail. Such information can be used to guide future research that can lead to the development of greatly enhanced representations for the rock stimuli. Figure 34 indicates that the models' underestimated performance for pumice in the averaged data is mostly due mis-predictions involving Pumice 4. Unlike the other examples of pumice used in this experiment, which were a light beige color, pumice 4 is a dark reddish-brown color. Thus, the models may have mis-categorized this rock because the CNN and MDS representations place a lot of emphasis on color, while people probably placed more emphasis on holes and other distinctive features. Similarly, the models overestimated performance for Granite 4. The models likely predicted high performance for this rock because it has similar colors (red and black) to Granite 1, one of the training exemplars. However, Granite 4 is a mostly black rock with red splotches, and Granite 1 is a mostly red rock with black splotches. This distinction is likely one that was salient to the human observers. Indeed, people likely confused granite and diorite with each other more often than the models predicted because people noticed that both

granite and diorite tend to be light or warm-colored rocks with dark spots, while the models did not have access to that information. Another interesting case where the models disagree with the data is Gneiss 3 which is layered. While the other examples of gneiss have stripes, they do not have physical layers, and while the models do not seem to make a distinction between the two, humans certainly do. All of these points again reinforce the idea that the CNN and MDS representations lack some important psychological dimensions to which people attend when learning the categories.

Discussion

Overall, both the MDS and CNN representations, when combined with the GCM, are able to make good first-order predictions of human categorization behavior. While the MDS representations provide somewhat better quantitative fits, the CNN representations make similar qualitative predictions, and in some cases are actually more accurate. Although the work is still in its early stages, the results show promise that deep learning may be used to automatically extract psychological representations from images, making more studies using natural categories feasible. The models' predictions become less accurate when the data is broken down into finer details, indicating that more research is needed to develop more accurate models of psychological representations and categorization. Possible directions this research may take are outlined in the General Discussion.

General Discussion

Summary and Implications

In this dissertation, I have shown that deep convolutional neural networks can be trained to predict the MDS coordinates derived from human similarity judgments of images of rocks. In addition, the CNNs can generalize to new rocks that were not even part of the MDS solution that the networks were trained on. Finally, the CNN-derived representations can be used in conjunction with the generalized context model to yield good first-order predictions of human categorization performance involving the rock stimuli. These results provide some promise that deep learning may be used to automate time- and resource-intensive MDS studies in the future, making it feasible for researchers to conduct more large-scale studies using natural stimuli.

There will likely always be a place for categorization research using artificial stimuli, since such stimuli allow researchers to conduct highly controlled experiments for testing between competing theoretical views. However, research using natural stimuli may also challenge certain theories. Recall from the introduction that the high between-category similarity and low within-category similarity found in Nosofsky et al.'s (2017) MDS solution demonstrates an important exception to the family-resemblance principle that has long been thought to govern natural categories (Rosch & Mervis, 1975). My findings suggest that deep learning could be used to derive feature-space representations of an even greater variety of rocks, beyond just the 30 subtypes in Nosofsky et al.'s (2017) data set. This would provide an even stronger test of the family-resemblance principle and may allow for new theories regarding the structure of natural categories to develop.

It should be straightforward to apply my deep learning procedure to other types of natural stimuli as well. Following Nosofsky et al. (2017), an initial MDS solution may be derived for a

representative sample of the stimuli by collecting similarity judgments and direct dimension ratings along hypothesized dimensions, and then CNNs can be trained on the MDS solution and used to generate representations of novel stimuli. One interesting domain on which to apply this procedure would be skin lesions, which also seem to violate the family-resemblance principle. Xu et al. (2016) argue that traditional rule-based approaches for teaching dermatology students to categorize skin lesions are ineffective because many benign and malignant examples share the same features. For example, irregular borders and variegated color are often indicators of cancer, but many benign lesions also show these features, and many malignant lesions in their early stages do not. A feature-space representation of this domain could be used in conjunction with GCM or other formal models to guide the search for more effective teaching techniques.

As I explain below, further research will be needed to evaluate how my technique compares to alternative methods for extracting psychological representations of stimuli, using deep learning or otherwise. Regardless of the outcome, though, I believe that this work should be of interest to many researchers, even those outside of the domain of categorization. While other researchers have found that networks may indirectly learn representations similar to those used by humans after being trained to perform other tasks, the present work is, to my knowledge, the first attempt to directly train networks to produce psychological dimensions. Furthermore, this work can be used as a tutorial for how psychologists can utilize the power of deep learning in novel ways to solve their research problems.

Directions for future research

All of the results reported in this dissertation should be regarded as a proof of concept and not the absolute best results that my technique could produce. I describe below various

directions of future research that may lead to better models of psychological representations and more accurate categorization predictions.

Improving the psychological feature space. The quality of the CNN predictions are dependent on the quality of the data they are trained on, so the CNN predictions may be improved by eliminating some of the shortcomings of Nosofsky et al.'s (2017) MDS solution. One issue with this MDS solution was that it was derived from a noisy similarity matrix with many entries based on very few observations, or no observations at all. Collecting more data to fill out the similarity matrix is one solution to this problem. Another solution may be to collect similarity judgments between pairs of rocks taken from Nosfosky et al.'s (2017) 360-rock data set and from the 120-rock data set reported in this dissertation, and then to create a shared 480 rock MDS solution. Increasing the number of items in the MDS space may impose stronger constraints on where each item can be located, resulting in more accurate similarity relationships. Furthermore, a bigger MDS solution would create more training data for the CNNs, which would further improve their predictive power.

Another limitation associated with the MDS solution is that it contains only 8 dimensions, but people probably attend to dozens of dimensions in these rock stimuli. This means that some of the MDS dimensions are probably amalgamations of several underlying psychological dimensions. This may explain why the "shape" dimension has eluded a more substantial interpretation, for example, but even some of the dimensions with seemingly clearer interpretations may be obscuring some important details. For instance, the "organization" dimension divides rocks into those with fragments and those with layers or stripes, but it does not differentiate rounded versus angular fragments or straight versus curved stripes, and these distinctions are important for classifying certain categories of rocks. And as mentioned

previously, there are important diagnostic dimensions that are missing from the MDS solution entirely, such as the presence of holes. Such features may not have been salient in the context of the generic similarity-judgment task used to derive the MDS solution, but may become highly salient in the context of a task in which subjects learn to classify the rocks into categories. Thus, it may be necessary to manually add these features to the psychological feature space to better model people's categorization performance. Nosofsky et al. (2017) outline other avenues for building a better psychological feature space.

Alternative techniques for automatically extracting psychological representations. The approach I took in this dissertation to automatically extracting psychological representations was to train CNNs to produce the MDS coordinates of the rock images, but as mentioned in the Related Work section, other researchers have taken different approaches, and it will be fruitful to compare these alternatives against each other. Peterson et al.'s (2016) approach of simply using the activations from the hidden layers of pre-trained CNNs as the feature space representations is especially attractive given that it does not require additional training. And while there is reason to be skeptical that the representations derived from these pre-trained CNNs will correspond to those used by humans, it is important to remember that the MDS dimensions do not represent the ground truth either, especially given the discussion above. Neither the CNN hidden-layer activations nor the MDS representations will be perfect models of psychological representations, but it will need to be determined if one is more useful than the other.

Given that there is much noise in Nosofsky et al.'s (2017) similarity-judgments data set, it may not be possible to directly apply Rumelhart and Todd's (1993) technique of training neural networks to produce similarity judgments between pairs of stimuli and then to extract the learned representations in the hidden layers. While some of this noise may be removed simply by

collecting more data, it may also be worthwhile to try a similar approach in which learned representations are extracted from networks directly trained to produce the category labels of rock images. Indeed, it would be interesting to see whether such networks would learn representations similar to those used by humans, and if those representations would include the category-specific dimensions that the MDS solution lacks, such as the presence of holes. A related question is whether the networks would make the same sorts of errors as humans. If CNNs do make the same errors as humans, that may provide evidence that they can be used to model the human visual system. And if CNNs do not make the same errors as humans, it may still be still be interesting to see whether it could be forced to by training it to produce human categorization judgments rather than the ground truth labels. In either case, the CNNs' ability to predict human categorization behavior could be compared against traditional cognitive models such as GCM. The best performing models could be used to help guide the search for effective teaching techniques and to thereby make recommendations to science-category educators

It is possible that the best network for deriving psychological representations would be one that is trained to simultaneously produce several different types of data. A single network could be trained, for example, to produce the MDS coordinates and category labels of individual rocks, as well as the similarity relationships between pairs of rocks. The intuition here is that providing the network with explicit similarity relationships and category information will constrain where the network can place each rock in the multidimensional space and force it to make more accurate predictions. Indeed, this sort of multi-task learning has been shown to improve the performance of neural networks in other tasks by causing the networks to learn robust shared representations across tasks (Caruana, 1998), so future research will need to explore this technique.

Deep learning is also not the only potential solution to efficiently deriving psychological representations. For example, in the spatial arrangement method, instead of collecting similarity ratings for individual pairs of items, participants are simply asked to arrange items on a computer screen such that similar items are close together and dissimilar items are far apart. A distance matrix can then be computed using the pixel coordinates of each item, and MDS analyses can be conducted in the usual way. This method takes much less time than traditional techniques for collecting similarity judgments, but it has been shown to yield similar MDS solutions for simple artificial stimuli (Goldstone, 1994; Hout, Goldinger, & Ferguson, 2013), and a preliminary study has shown that it may also scale up to the rocks stimuli (S. Goldfinger, M. Hout, R. Nosofsky, personal communication, April 26, 2018). It should also be explored whether meaningful dimensions could be derived by applying MDS or other dimension-reduction techniques on the raw pixel values of images.

Improving the generalized context model. While the discussion thus far has focused on the shortcomings of the psychological feature space, there are also undoubtedly shortcomings in the generalized context model. For example, the full version of the GCM model presented in this dissertation made use of 50 free parameters, and it would be more satisfying if the values of these parameters could be derived from theory rather than simply fit to the data. For example, research has shown that both familiarity and novelty influence people's preferences for choosing some visual categories over others (Park, Shimojo, & Shimojo, 2010), so the category response bias parameters in GCM could potentially be estimated by measuring people's prior knowledge or familiarity with each category of rock. And since it is often assumed that people will allocate their attention to optimize categorization, attention weight parameters that maximize categorization accuracy could be estimated as in previous applications of GCM (Nosofsky, 1984,

1986). It may also be informative to compute attention weight values according to various machine learning algorithms that rank the relative importance of dimensions (Guyon & Elisseeff, 2003).

Ideas from machine learning may also be useful for rethinking how GCM uses its attention weight parameters in more fundamental ways. The implementation of GCM described in this dissertation evaluates the evidence that a stimulus belongs in each category by simultaneously comparing the stimulus to all exemplars in all categories, using a single set of attention weights. However, people might actually make several different types of comparisons using different sets of attention weights before making their final categorization decision. For instance, a person may first evaluate the evidence that the rock belongs to obsidian by comparing it to all exemplars of obsidian and all exemplars of not-obsidian, paying special attention to how dark and shiny the rock is. Then the person may evaluate the evidence that the rock belongs to pumice by comparing it to all exemplars of pumice and all exemplars of not-pumice, paying special attention to whether it has holes, etc. This type of model is known as a one-vs-all classifier in the machine learning literature (Bishop, 2006). In this type of scheme, a separate classifier is trained for each category. Each classifier produces a confidence score that the stimulus belongs to the classifier's associated category, and the final categorization decision is made according to whichever category has the highest confidence score.

Alternatively, people may compare the relative evidence between pairs of categories. A person may first compare the relative evidence between obsidian and anthracite, paying special attention to whether the surfaces are scalloped or bumpy. The person then may compare the evidence between conglomerate and breccia, paying special attention to whether the fragments are round or smooth, etc. This type of model is known as a one-vs-one classifier in the machine

learning literature, (Bishop, 2006). In this type of scheme, with *K* categories, *K(K-1)/2* separate classifiers are trained. Each classifier produces confidence scores for pairs of categories and votes for whichever category has the higher confidence score. The final categorization decision is made according to whichever category received the most total votes. Future research should explore whether one-vs-all or a one-vs-one versions of GCM can better capture human categorization behavior.

It is also important to keep in mind that GCM and the entire exemplar framework is just one of many competing models of categorization. Other types of models include prototype models (Posner & Keele, 1968; J. D. Smith & Minda, 1998), decision bound models (e.g., Ashby & Maddox, 1993; Mckinley & Nosofsky, 1995) and clustering models (e.g., Anderson, 1991; Love, Medin, & Gureckis, 2004). Rule-plus-exception models such as RULEX (Nosofsky, 1992) or ATRIUM (Erickson & Kruschke, 1998) seem especially attractive given the nature of certain category distinctions. For example, people may remember rules such as "if a rock has holes, it must be pumice, unless it is also dark, in which case it is probably basalt." It could be the case that any of these alternative models come closer to describing the nature of people's category representations and decision processes and would yield better categorization predictions than the GCM.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., … Devin, M. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *ArXiv Preprint ArXiv:1603.04467*.

Agrawal, P., Stansbury, D., Malik, J., & Gallant, J. L. (2014). Pixels to Voxels: Modeling Visual Representation in the Human Brain. *ArXiv:1407.5104 [Cs, q-Bio]*. Retrieved from http://arxiv.org/abs/1407.5104

Akbas, E., & Eckstein, M. P. (2017). Object detection through search with a foveated visual system. *PLOS Computational Biology*, *13*(10), e1005743. https://doi.org/10.1371/journal.pcbi.1005743

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, *98*(3), 409.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400.

Bambach, S., Crandall, D. J., Smith, L. B., & Yu, C. (2016). Active viewing in toddlers facilitates visual object learning: An egocentric vision approach. In *Proceedings of the 38th Annual Conference of the Cognitive Science Society. Philadelphia, PA* (pp. 1631–1636).

Bambach, S., Zhang, Z., Crandall, D. J., & Yu, C. (2017). Exploring Inter-Observer Differences in First-Person Object Views using Deep Learning Models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2773–2782).

Battleday, R. M., Peterson, J. C., & Griffiths, T. L. (2017). Modeling Human Categorization of Natural Images Using Deep Feature Representations. *ArXiv:1711.04855 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1711.04855

Bengio, Y. (2012). Practical recommendations for gradient-based training of deep architectures. Retrieved from https://arxiv.org/abs/1206.5533

Bishop, C. M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, *10*, 433–436.

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., … DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*, *10*(12), e1003963. https://doi.org/10.1371/journal.pcbi.1003963

Carlini, N., & Wagner, D. (2016). Towards Evaluating the Robustness of Neural Networks. *ArXiv:1608.04644 [Cs]*. Retrieved from http://arxiv.org/abs/1608.04644

Caruana, R. (1998). Multitask Learning. In *Learning to Learn* (pp. 95–133). Springer, Boston, MA. https://doi.org/10.1007/978-1-4615-5529-2_5

Chollet, F., & others. (2015). *Keras*. Github. Retrieved from https://github.com/keras-team/keras

Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*(6203), 129–132.

Cybenko, G. (1989). Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, *2*(4), 303–314. https://doi.org/10.1007/BF02551274

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition (pp. 647–655). Presented at the International conference on machine learning.

Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but Not Deep Neural Networks, Often Miss Giant Targets in Scenes. *Current Biology*, *27*(18), 2827-2832.e3. https://doi.org/10.1016/j.cub.2017.07.068

Elsayed, G. F., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial Examples that Fool both Human and Computer Vision. *ArXiv:1802.08195 [Cs, q-Bio, Stat]*. Retrieved from http://arxiv.org/abs/1802.08195

Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology-General*, *127*(2), 107–140. https://doi.org/10.1037//0096-3445.127.2.107

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *ArXiv Preprint ArXiv:1706.06969*.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers*, *26*(4), 381–386. https://doi.org/10.3758/BF03204653

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. *ArXiv:1412.6572 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1412.6572

Gower, J. C., & Dijksterhuis, G. B. (2004). *Procrustes Problems*. OUP Oxford.

Gu, S., & Rigazio, L. (2014). Towards Deep Neural Network Architectures Robust to Adversarial Examples. *ArXiv:1412.5068 [Cs]*. Retrieved from http://arxiv.org/abs/1412.5068

Guyon, I., & Elisseeff, A. (2003). An Introduction to Variable and Feature Selection. *J. Mach. Learn. Res.*, *3*, 1157–1182.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(10), 993–1001. https://doi.org/10.1109/34.58871

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition (pp. 770–778). Presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Hout, M. C., Goldinger, S. D., & Ferguson, R. W. (2013). The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General*, *142*(1), 256–281. https://doi.org/10.1037/a0028860

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift (pp. 448–456). Presented at the International Conference on Machine Learning.

JASP Team. (2018). *JASP (Version 0.8.6)[Computer software]*. Retrieved from https://jasp-stats.org/

Kass, R. E., & Raftery, A. E. (1995). Bayes Factors. *Jou 10rnal of the American Statistical Association*, *90*(430), 773–795. https://doi.org/10.1080/01621459.1995.10476572

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, *10*(11), e1003915. https://doi.org/10.1371/journal.pcbi.1003915

Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. *Scientific Reports*, *6*, 32672. https://doi.org/10.1038/srep32672

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks (pp. 1097–1105). Presented at the Advances in neural information processing systems.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, *29*(1), 1–27.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling* (Vol. 11). Sage.

Kümmerer, M., Theis, L., & Bethge, M. (2014). Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. *ArXiv:1411.1045 [Cs, q-Bio, Stat]*. Retrieved from http://arxiv.org/abs/1411.1045

Kümmerer, M., Wallis, T., & Bethge, M. (2017). DeepGaze II: Predicting fixations from deep features over time and tasks. *Journal of Vision*, *17*(10), 1147–1147. https://doi.org/10.1167/17.10.1147

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. *ArXiv:1607.02533 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1607.02533

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015a). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015b). Deep Neural Networks Predict Category Typicality Ratings for Images. Presented at the CogSci.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*(4), 541–551.

Lee, M. D. (2001). Determining the Dimensionality of Multidimensional Scaling Representations for Cognitive Modeling. *Journal of Mathematical Psychology*, *45*(1), 149–166. https://doi.org/10.1006/jmps.1999.1300

Lee, S., Purushwalkam, S., Cogswell, M., Crandall, D., & Batra, D. (2015). Why M Heads are Better than One: Training a Diverse Ensemble of Deep Networks. *ArXiv:1511.06314 [Cs]*. Retrieved from http://arxiv.org/abs/1511.06314

Liu, W., Dai, B., Humayun, A., Tay, C., Yu, C., Smith, L. B., … Song, L. (2017). Iterative Machine Teaching. *ArXiv:1705.10470 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1705.10470

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: a network model of category learning. *Psychological Review*, *111*(2), 309.

Maddox, W. T., Ashby, F. G., & Bohil, C. J. (2003). Delayed feedback effects on rule-based and information-integration category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 650.

Mathy, F., & Feldman, J. (2016). The influence of presentation order on category transfer. *Experimental Psychology*, *63*(1), 59–69. https://doi.org/10.1027/1618-3169/a000312

Mckinley, S. C., & Nosofsky, R. M. (1995). Investigations of Exemplar and Decision Bound Models in Large, Ill-Defined Category Structures. *Journal of Experimental Psychology-Human Perception and Performance*, *21*(1), 128–148. https://doi.org/10.1037/0096-1523.21.1.128

Meagher, B. J., Catalado, K., Douglas, B. J., & Nosofsky, R. M. (submitted). Training of rock classifications: The use of computer images versus physical-rock samples. *Journal of Geoscience Education*.

Nair, V., & Hinton, G. E. (2010). Rectified linear units improve restricted boltzmann machines (pp. 807–814). Presented at the Proceedings of the 27th international conference on machine learning (ICML-10).

Nguyen, A., Yosinski, J., & Clune, J. (2014). Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. *ArXiv:1412.1897 [Cs]*. Retrieved from http://arxiv.org/abs/1412.1897

Nosofsky, R. M. (1984). Choice, Similarity, and the Context Theory of Classification. *Journal of Experimental Psychology-Learning Memory and Cognition*, *10*(1), 104–114. https://doi.org/10.1037/0278-7393.10.1.104

Nosofsky, R. M. (1986). Attention, Similarity, and the Identification-Categorization Relationship. *Journal of Experimental Psychology-General*, *115*(1), 39–57. https://doi.org/10.1037/0096-3445.115.1.39

Nosofsky, R. M. (1992). Rule-Plus-Exception Model of Classification Learning. *Bulletin of the Psychonomic Society*, *30*(6), 448–448.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. *Formal Approaches in Categorization*, 18–39.

Nosofsky, R. M., Sanders, C. A., Gerdom, A., Douglas, B. J., & McDaniel, M. A. (2017). On Learning Natural-Science Categories That Violate the Family-Resemblance Principle. *Psychological Science*, *28*(1), 104–114. https://doi.org/10.1177/0956797616675636

Nosofsky, R. M., Sanders, C. A., & McDaniel, M. A. (2018). Tests of an Exemplar-Memory Model of Classification Learning in a High-Dimensional Natural-Science Category Domain.

Nosofsky, R. M., Sanders, C. A., Meagher, B. J., & Douglas, B. J. (2017). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods*. https://doi.org/10.3758/s13428-017-0884-8

Nosofsky, R. M., Sanders, C. A., Zhu, X., & McDaniel, M. A. (submitted). Model-Guided Search for Optimal Training Exemplars in a Natural-Science Category Domain.

Olshausen, B. A. (2013). 20 Years of Learning About Vision: Questions Answered, Questions Unanswered, and Questions Not Yet Asked. In *20 Years of Computational Neuroscience* (pp. 243–270). Springer, New York, NY. https://doi.org/10.1007/978-1-4614-1424-7_12

Papernot, N., McDaniel, P., & Goodfellow, I. (2016). Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. *ArXiv:1605.07277 [Cs]*. Retrieved from http://arxiv.org/abs/1605.07277

Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks. In *2016 IEEE Symposium on Security and Privacy (SP)* (pp. 582–597). https://doi.org/10.1109/SP.2016.41

Park, J., Shimojo, E., & Shimojo, S. (2010). Roles of familiarity and novelty in visual preference judgments are segregated across object categories. *Proceedings of the National Academy of Sciences*, *107*(33), 14552–14555. https://doi.org/10.1073/pnas.1004374107

Patil, K. R., Zhu, X., Kopeć, Ł., & Love, B. C. (2014). Optimal teaching for limited-capacity human learners (pp. 2465–2473). Presented at the Advances in Neural Information Processing Systems.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*(Oct), 2825–2830.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2016). Adapting Deep Network Features to Capture Psychological Representations. *ArXiv:1608.02164 [Cs]*. Retrieved from http://arxiv.org/abs/1608.02164

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2017). Adapting deep network features to capture psychological representations: an abridged report. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence* (pp. 4934–4938). AAAI Press.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3p1), 353.

Pothos, E. M., & Wills, A. J. (2011). *Formal approaches in categorization*. Cambridge University Press.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *BioRxiv*, 240614. https://doi.org/10.1101/240614

Richler, J. J., & Palmeri, T. J. (2014). Visual category learning. *Wiley Interdisciplinary Reviews: Cognitive Science*, *5*(1), 75–94. https://doi.org/10.1002/wcs.1268

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025. https://doi.org/10.1038/14819

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605. https://doi.org/10.1016/0010-0285(75)90024-9

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology*, *53*(2), 94–101. https://doi.org/10.1037/h0041867

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error-propagation. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1* (Vol. 1, pp. 318–362). MIT Press, Cambridge, MA.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*, 3–30.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

Sharif Razavian, A., Azizpour, H., Sullivan, J., & Carlsson, S. (2014). CNN features off-the-shelf: an astounding baseline for recognition (pp. 806–813). Presented at the Proceedings of the IEEE conference on computer vision and pattern recognition workshops.

Shepard, R. N. (1963). Analysis of Proximities as a Technique for the Study of Information Processing in Man. *Human Factors*, *5*(1), 33–48. https://doi.org/10.1177/001872086300500104

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*.

Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411.

Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The Developing Infant Creates a Curriculum for Statistical Learning. *Trends in Cognitive Sciences*, *22*(4), 325–336. https://doi.org/10.1016/j.tics.2018.02.004

Springenberg, J. T., Dosovitskiy, A., Brox, T., & Riedmiller, M. (2014). Striving for Simplicity: The All Convolutional Net. *ArXiv:1412.6806 [Cs]*. Retrieved from http://arxiv.org/abs/1412.6806

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, *15*(1), 1929–1958.

Stork, D. G. (1989). Is backpropagation biologically plausible? In *International 1989 Joint Conference on Neural Networks* (pp. 241–246 vol.2). https://doi.org/10.1109/IJCNN.1989.118705

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision (pp. 2818–2826). Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *ArXiv:1312.6199 [Cs]*. Retrieved from http://arxiv.org/abs/1312.6199

Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on* (pp. 3156–3164). IEEE.

Xu, B., Rourke, L., Robinson, J. K., & Tanaka, J. W. (2016). Training Melanoma Detection in Photographs Using the Perceptual Expertise Training Approach. *Applied Cognitive Psychology*, *30*(5), 750–756. https://doi.org/10.1002/acp.3250

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks? (pp. 3320–3328). Presented at the Advances in neural information processing systems.

Tables

Table 1: Subtypes of igneous, metamorphic, and sedimentary rocks used in (Nosofsky, Sanders, Meagher, et al., 2017) and the present work.

Igneous	Metamorphic	Sedimentary
Andesite	Amphibolite	Bituminous Coal
Basalt	Anthracite	Breccia
Diorite	Gneiss	Chert
Gabbro	Hornfels	Conglomerate
Granite	Marble	Dolomite
Obsidian	Migmatite	Micrite
Pegmatite	Phyllite	Rock Gypsum
Peridotite	Quartzite	Rock Salt
Pumice	Schist	Sandstone
Rhyolite	Slate	Shale

Table 2: Model Fits to Validation Set

Model	MSE
Null	6.67
Scratch	1.856
Feature extraction	2.132
Transfer learning	1.494
Fine-tuned	1.330
Ensemble	1.298

Table 3: Experiment 1: Correlations between dimensions 1-6 of the rotated MDS solution and the hypothesized dimensions from the direct dimension-ratings experiment, along with each dimension's anchor labels.

Dimension	**Correlation**	**Anchor Labels**
1. Lightness of Color	.921	Darkest/ Medium/ Lightest
2. Average Grain Size	.794	No Visible Grain/ Medium/ Very Coarse
3. Roughness	.542	Smoothest/ Medium/ Roughest
4. Shininess	.858	Dullest/ Medium/ Shiniest
5. Organization	.570	Disorganized/ Medium/ Organized
6. Chromaticity	.798	No Color/ Cool Color/ Warmest Color

Table 4: Subtypes of the three higher-level categories of rocks used in the mixed condition.

Igneous	**Metamorphic**	**Sedimentary**
Basalt	Anthracite	Dolomite
Diorite	Marble	Micrite
Obsidian		Rock Gypsum
Pumice		Sandstone

Table 5: Number of free parameters of each version of GCM and its best-fitting negative log-likelihood and BIC score using the CNN- and MDS-derived representations,

		CNN Representations		**MDS Representations**	
Model	**Free parameters**	**-ln(L)**	**BIC**	**-ln(L)**	**BIC**
Baseline	2	4162	8344	3314	6647
Baseline + Weights	23	3777	7781	2922	6072
Baseline + Biases	29	3854	7996	3126	6538
Full	50	3443	7380	2768	6032
Baseline , $c_w = c_b$	1	5772	11553	4502	9015
Baseline + Weights, $c_w = c_b$	22	5393	11002	4152	8522
Baseline + Biases, $c_w = c_b$	28	5190	10657	3191	8658
Full, $c_w = c_b$	49	4787	10059	3815	8115

Table 6: Best-fit Full GCM parameter values using both the CNN- and MDS-derived representations

	CNN Representations			MDS Representations		
	Igneous Condition	**Metamorphic Condition**	**Sedimentary Condition**	**Igneous Condition**	**Metamorphic Condition**	**Sedimentary Condition**
c_w	1.547	1.547	1.547	2.355	2.355	2.355
c_b	2.58	2.58	2.58	3.508	3.508	3.508
b_1	0.056	0.053	0.094	0.076	0.084	0.078
b_2	0.067	0.179	0.265	0.094	0.120	0.154
b_3	0.112	0.079	0.098	0.080	0.098	0.102
b_4	0.062	0.106	0.167	0.087	0.089	0.116
b_5	0.156	0.154	0.073	0.082	0.087	0.063
b_6	0.099	0.080	0.056	0.104	0.099	0.067
b_7	0.071	0.092	0.060	0.105	0.104	0.152
b_8	0.101	0.076	0.050	0.185	0.119	0.086
b_9	0.171	0.069	0.103	0.122	0.077	0.114
b_{10}	0.104	0.111	0.034	0.063	0.125	0.068
w_1	0.151	0.098	0.069	0.237	0.107	0.136
w_2	0.097	0.144	0.091	0.173	0.156	0.138
w_3	0.283	0.055	0.175	0.040	0.127	0.062
w_4	0.170	0.209	0.177	0.203	0.238	0.191
w_5	0.098	0.119	0.192	0.078	0.066	0.124
w_6	0.055	0.028	0.080	0.076	0.033	0.065
w_7	0.000	0.204	0.206	0.133	0.120	0.174
w_8	0.146	0.143	0.010	0.060	0.153	0.111

Figures

Figure 1: Three examples of obsidian, anthracite, and rhyolite. Obsidian and anthracite exhibit high between-category similarity, whereas rhyolite exhibits low within-category similarity.

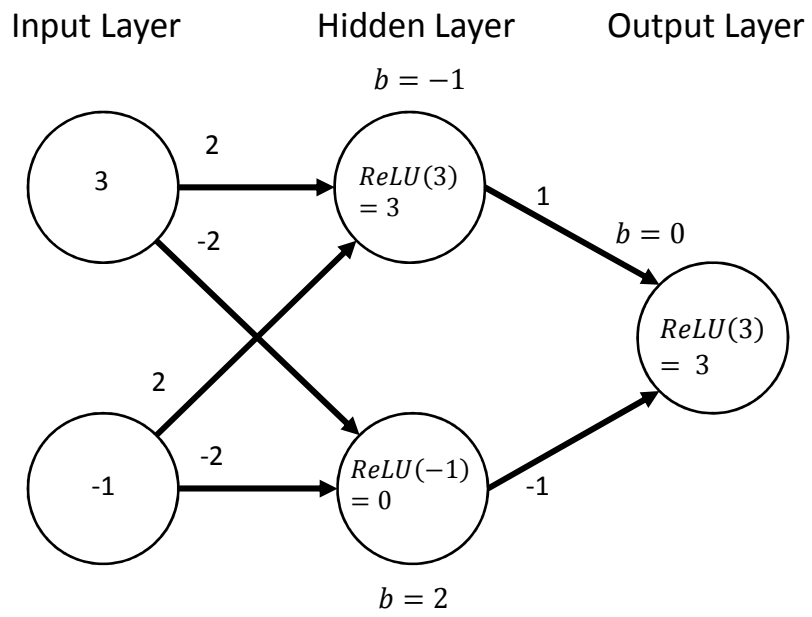

Figure 2: Multilayer perceptron. Numbers inside circles indicate unit activations, numbers above arrows indicate connection weights, and *b* values indicate unit biases.

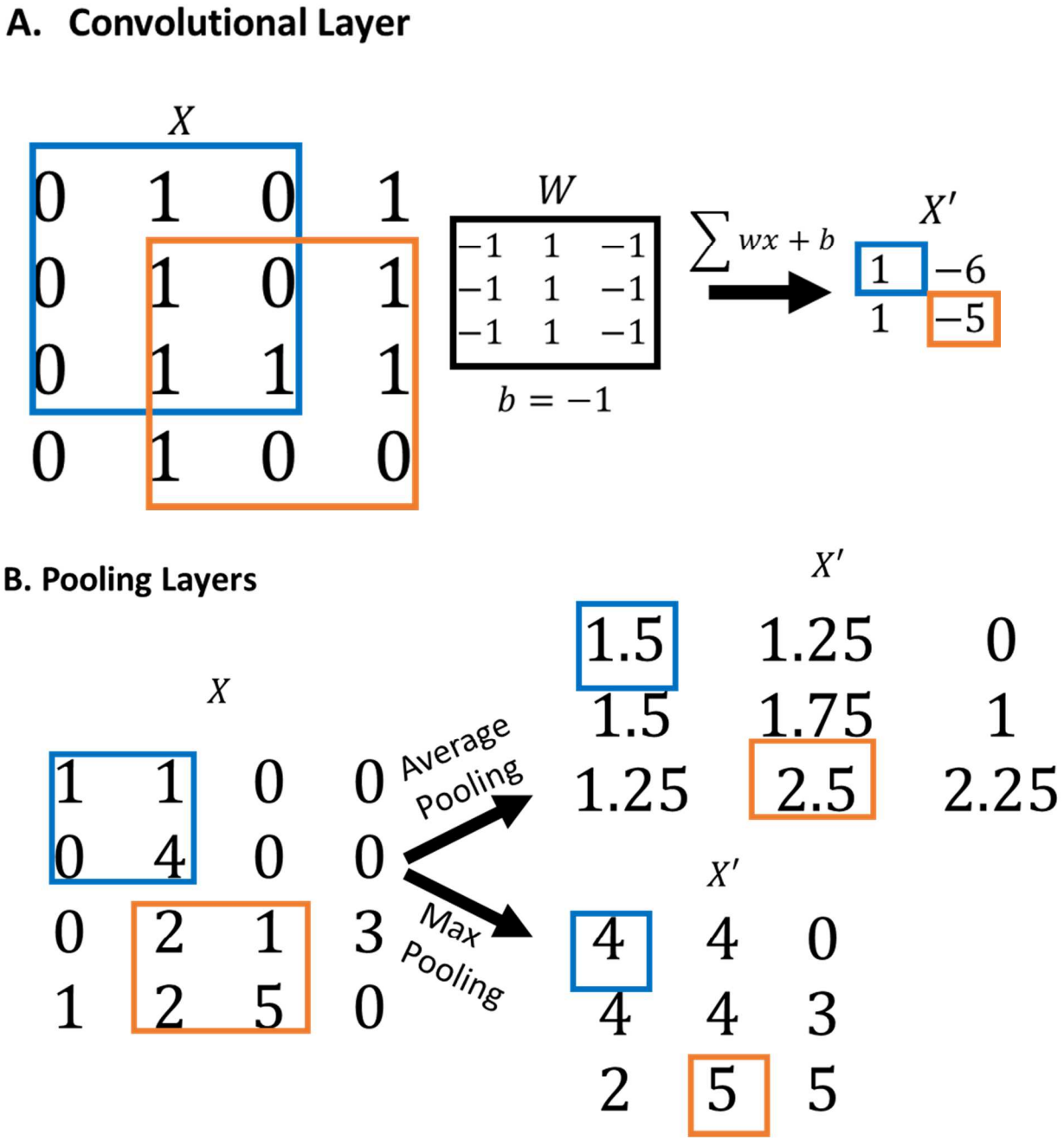

Figure 3: Convolutional and pooling layers in a convolutional neural network

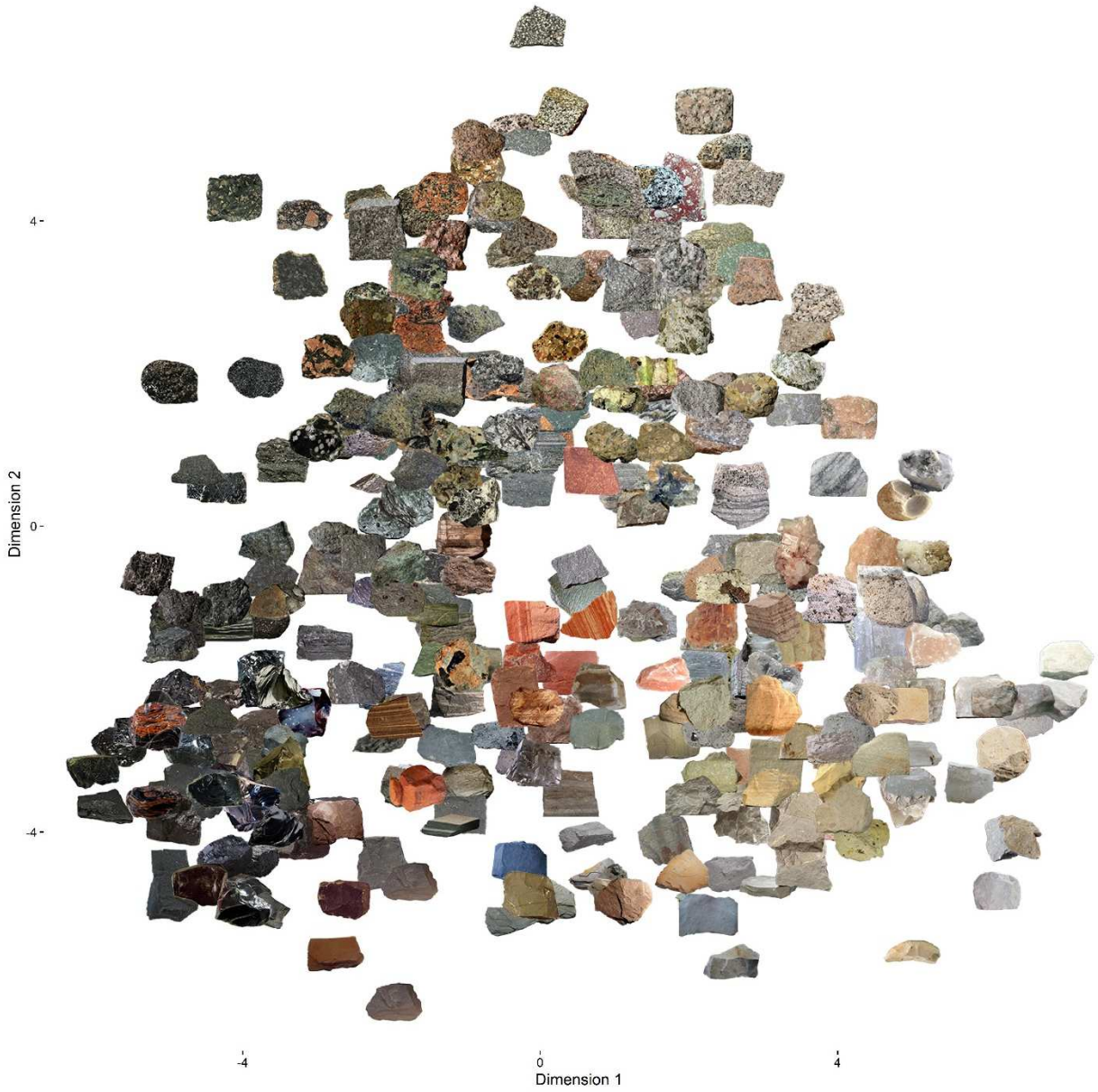

Figure 4: Plot of the first two rotated MDS dimensions from (Nosofsky, Sanders, Meagher, et al., 2017). Dimension 1 can be interpreted as the rocks' lightness/darkness, and dimension 2 can be interpreted as the rocks' average grain size.

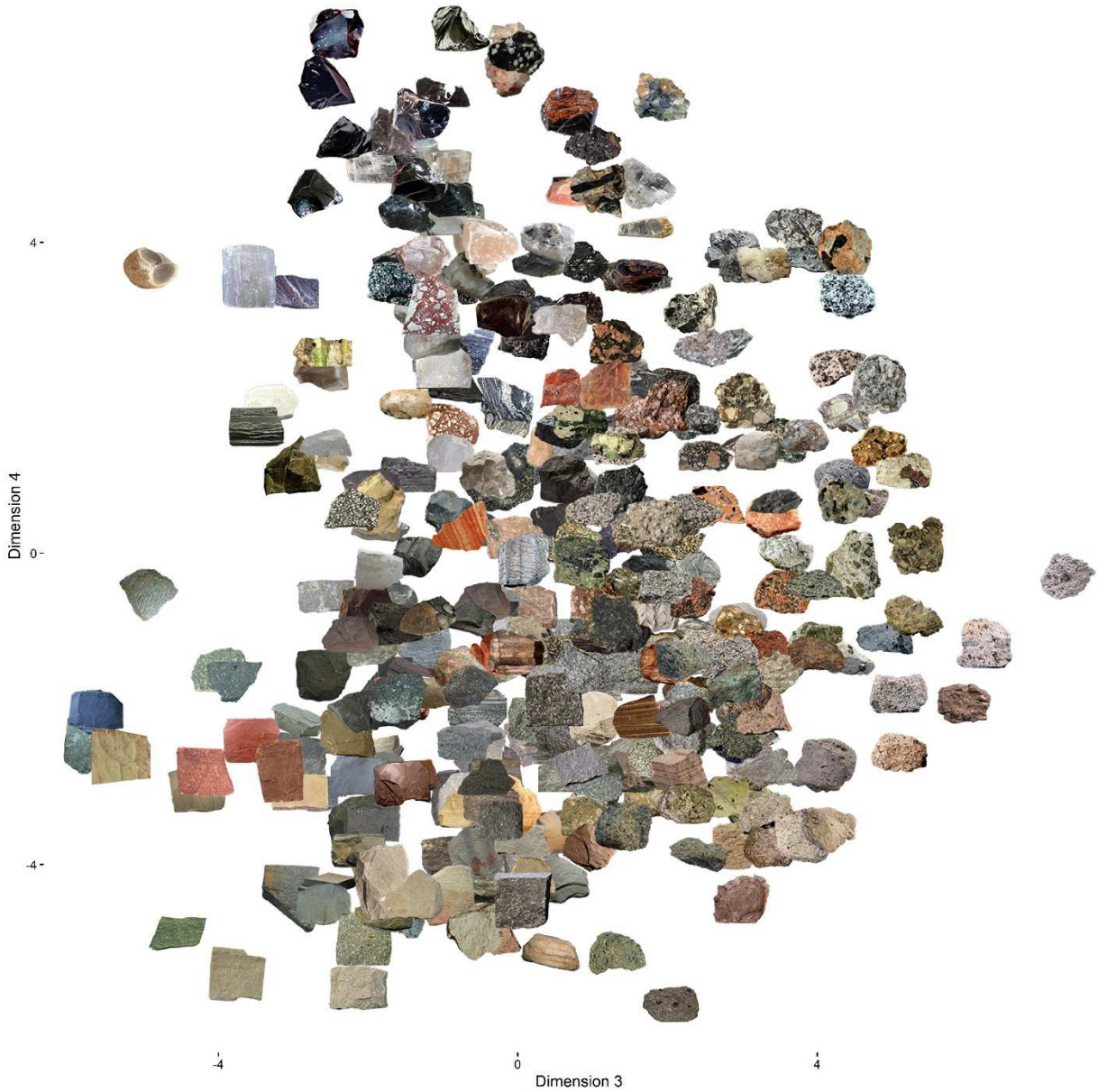

Figure 5: Plot of the third and fourth rotated MDS dimensions from (Nosofsky, Sanders, Meagher, et al., 2017). Dimension 3 can be interpreted as the rocks' roughness, and dimension 4 can be interpreted as the rocks' shininess.

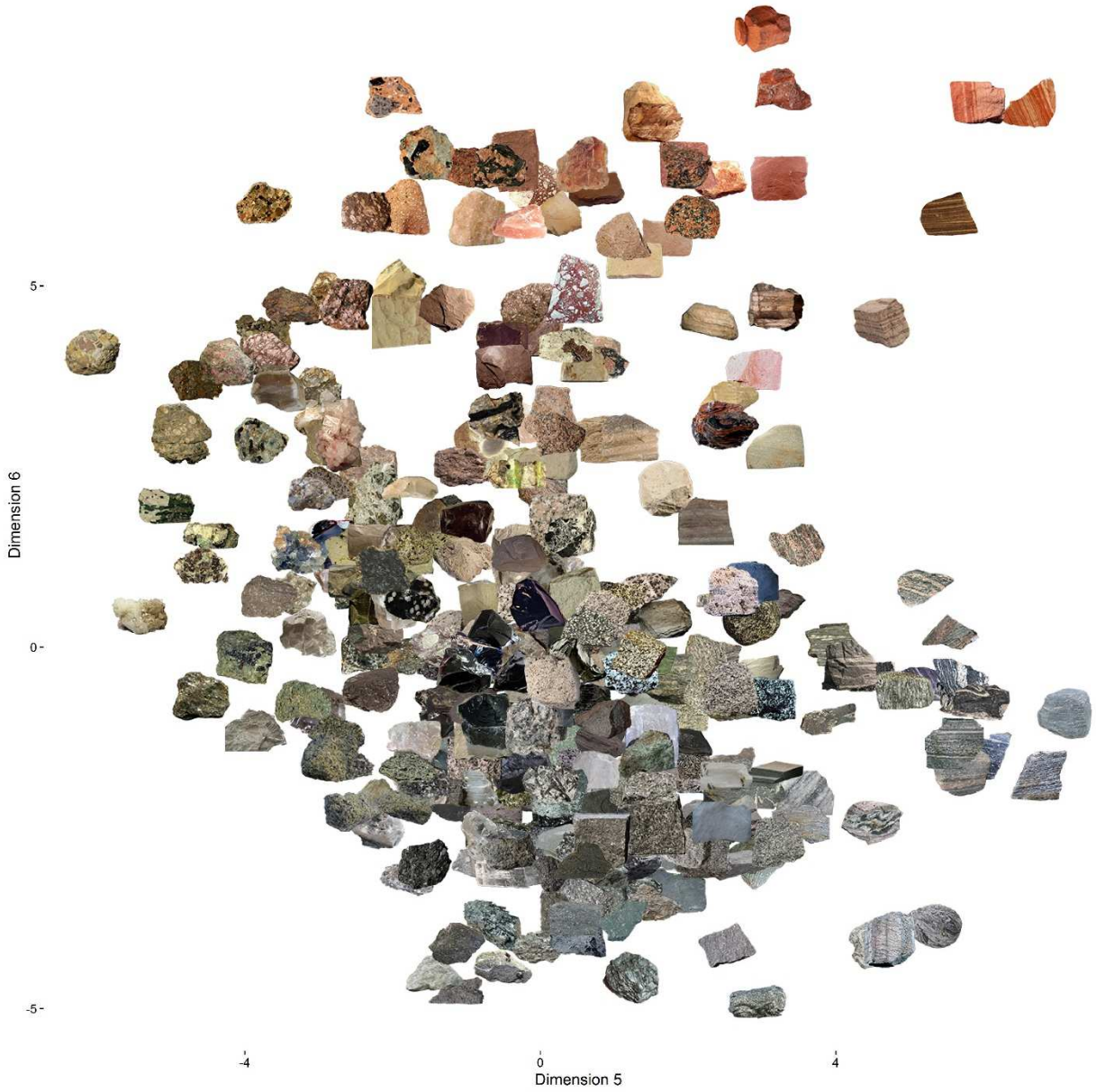

Figure 6: Plot of the fifth and sixth rotated MDS dimensions from (Nosofsky, Sanders, Meagher, et al., 2017). Dimension 5 can be interpreted as the rocks' organization (the extent to which a rock has organized layers or stripes vs. fragments haphazardly glued together), and dimension 6 can be interpreted as the rocks' chromaticity (the extent to which a rock's color is saturated/warm or desaturated/cool).

Figure 7: Plot of the seventh and eight rotated MDS dimensions from (Nosofsky, Sanders, Meagher, et al., 2017). While these dimensions do not have clear interpretations, dimension 7 seems to have some correspondence with the rocks' shape (rocks on the left side of the space tend to be flat, while rocks on the right tend to have more volume), and dimension 8 seems to have some correspondence with the rocks' hue (rocks on the top of the space tend to be green, while rocks on the bottom tend to be red).

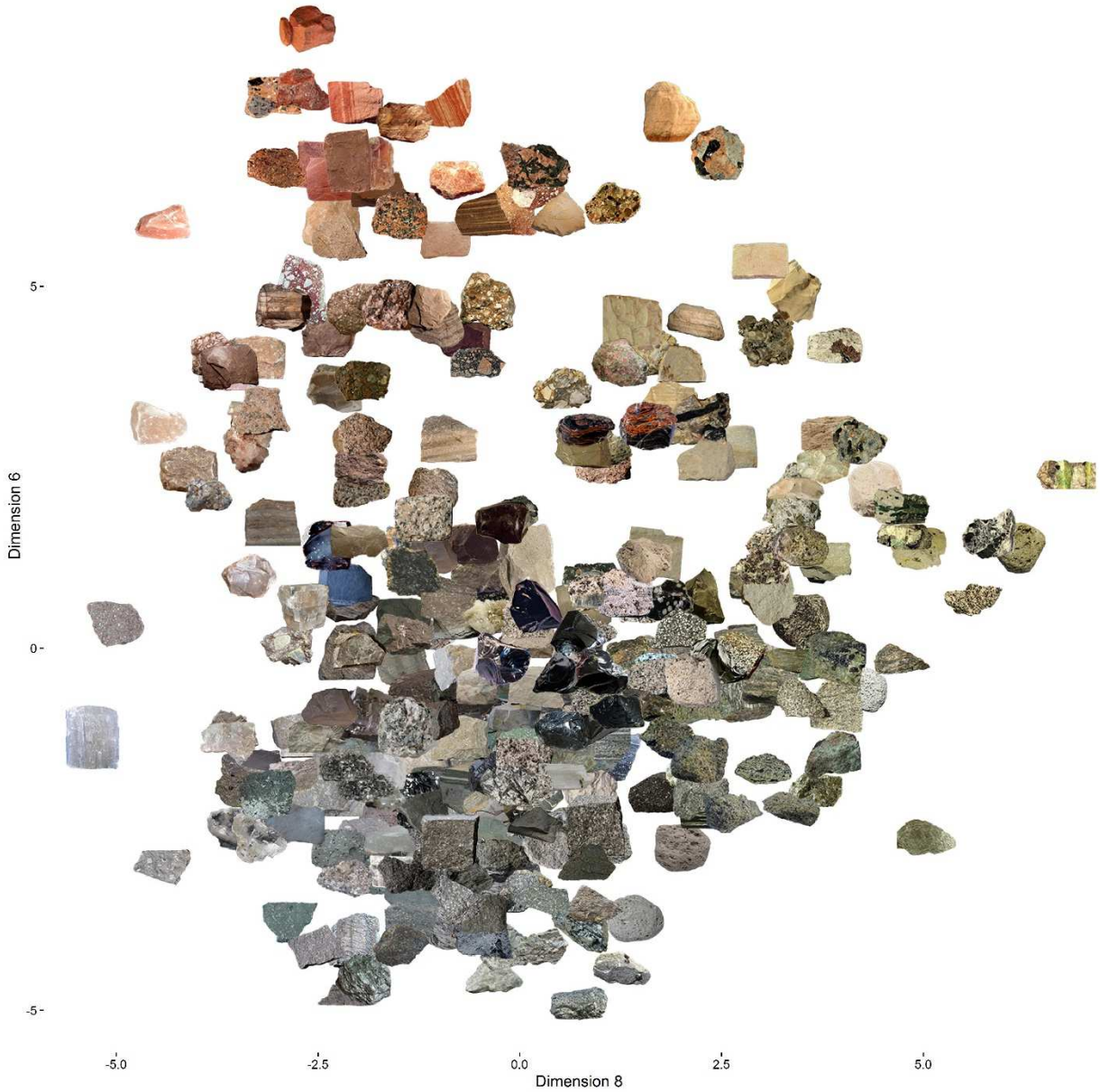

Figure 8: Plot of the eighth and sixth rotated MDS dimensions from (Nosofsky, Sanders, Meagher, et al., 2017). Plotting these dimensions together forms a loose color circle. Starting from the top and moving clockwise, the color of the rocks shifts from red, to orange, to yellow, to green, to blue, to violet, and then finally back to red.

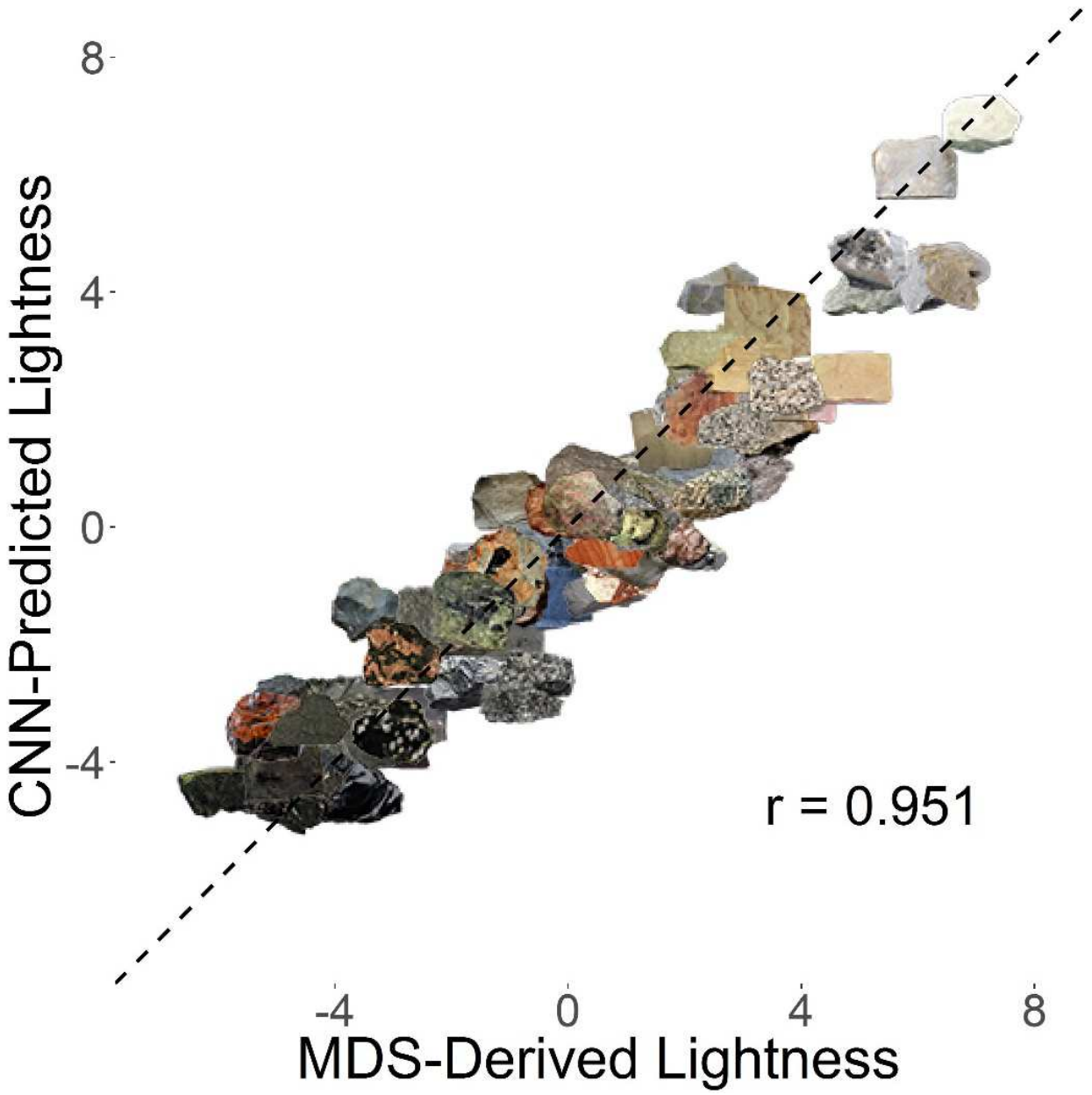

Figure 9: Scatterplot of Ensemble-predicted lightness against MDS-derived lightness for the test set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity.

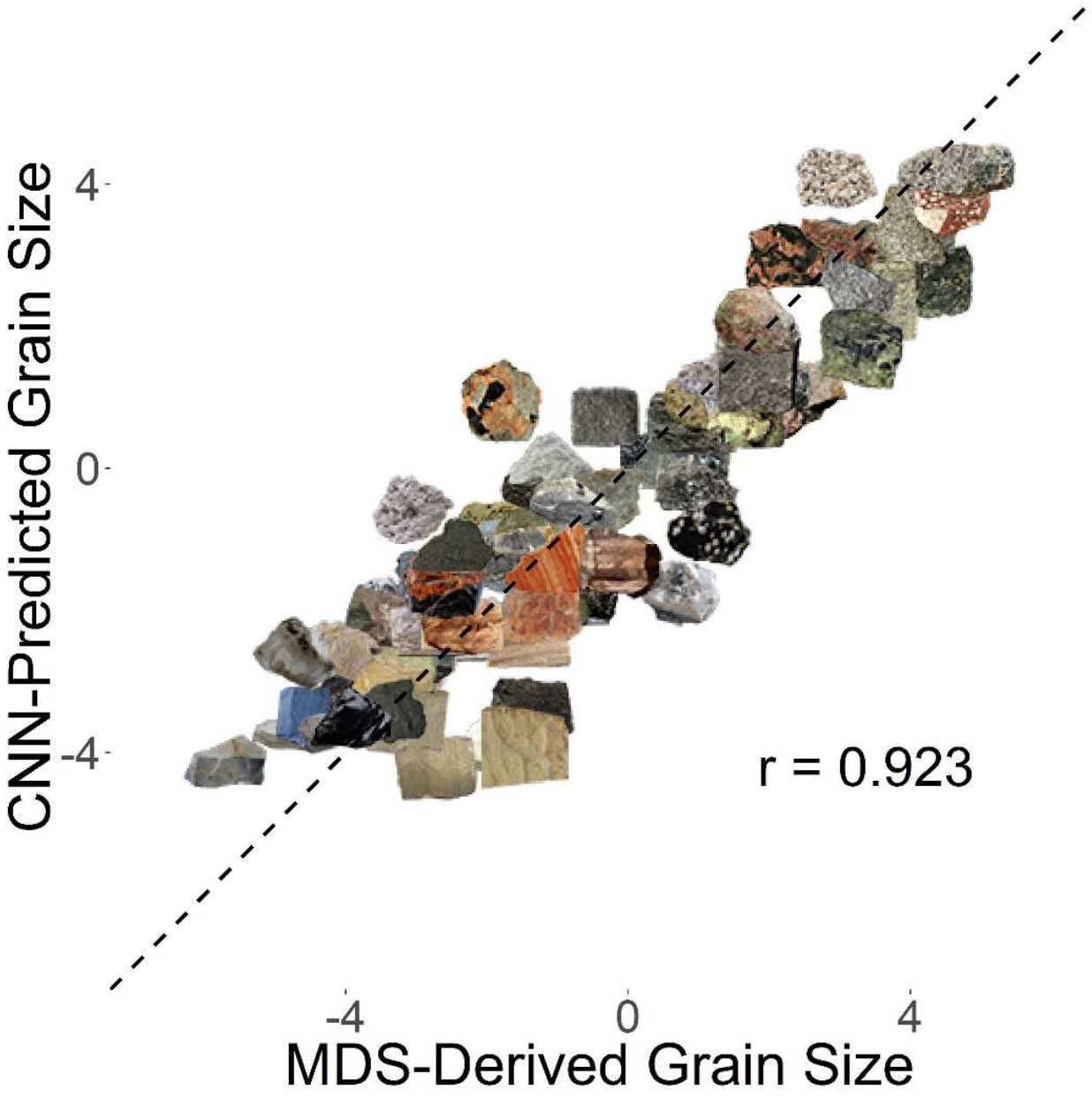

Figure 10: Scatterplot of Ensemble-predicted grain size against MDS-derived grain size for the test set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity.

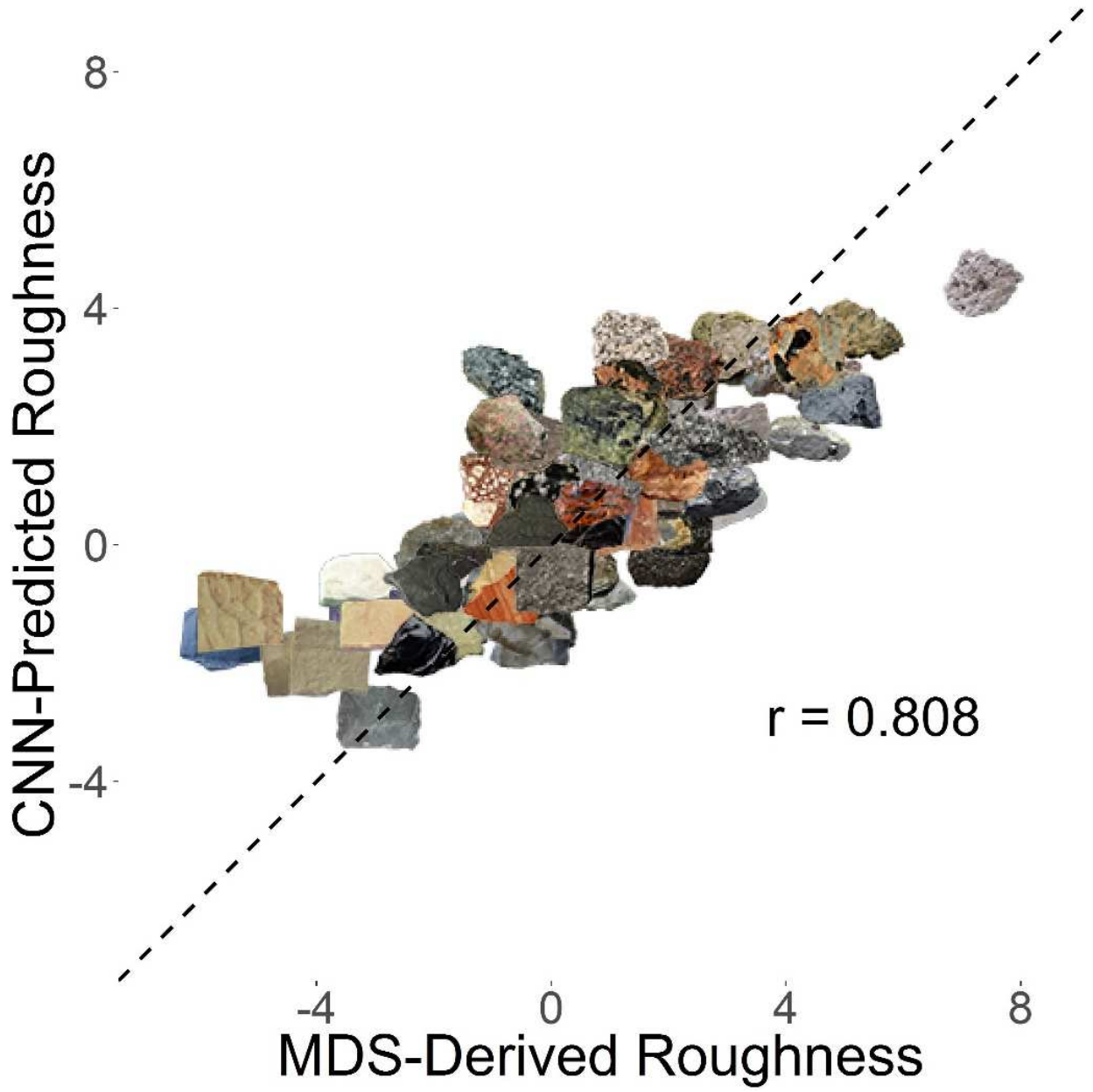

Figure 11: Scatterplot of Ensemble-predicted roughness against MDS-derived roughness for the test set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity.

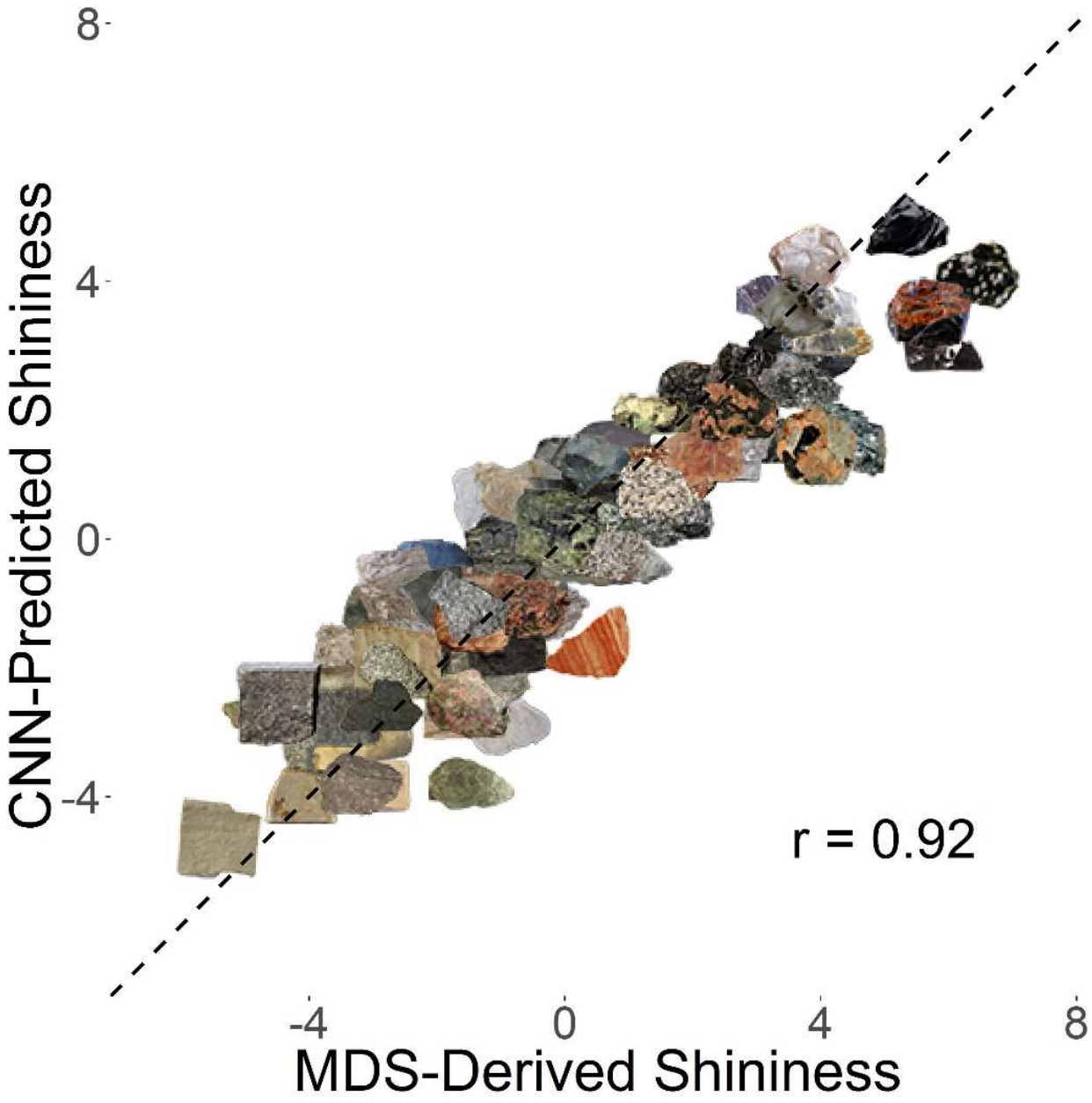

Figure 12: Scatterplot of Ensemble-predicted shininess against MDS-derived grain size for the test set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity.

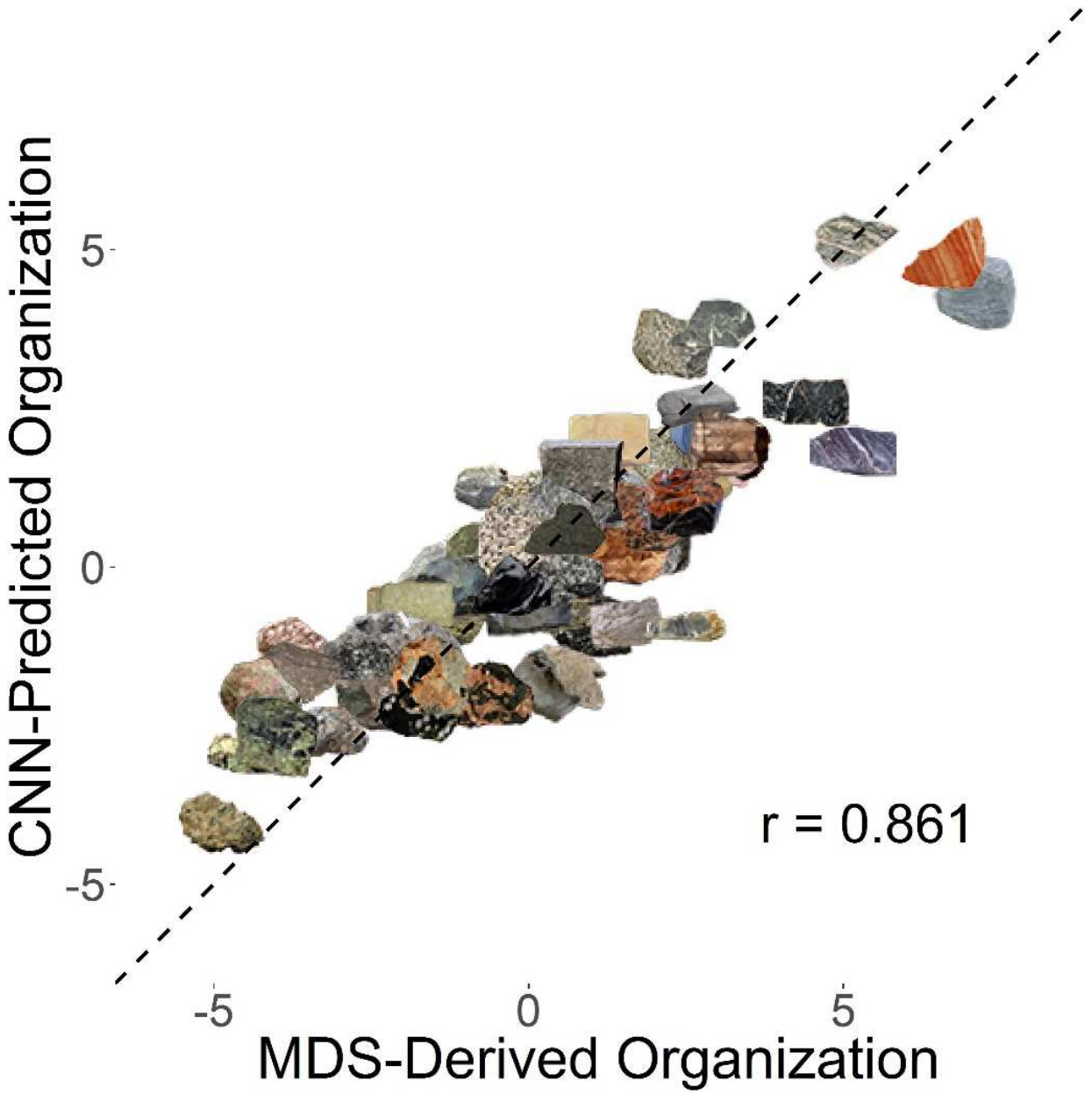

Figure 13: Scatterplot of Ensemble-predicted organization against MDS-derived organization for the test set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity.

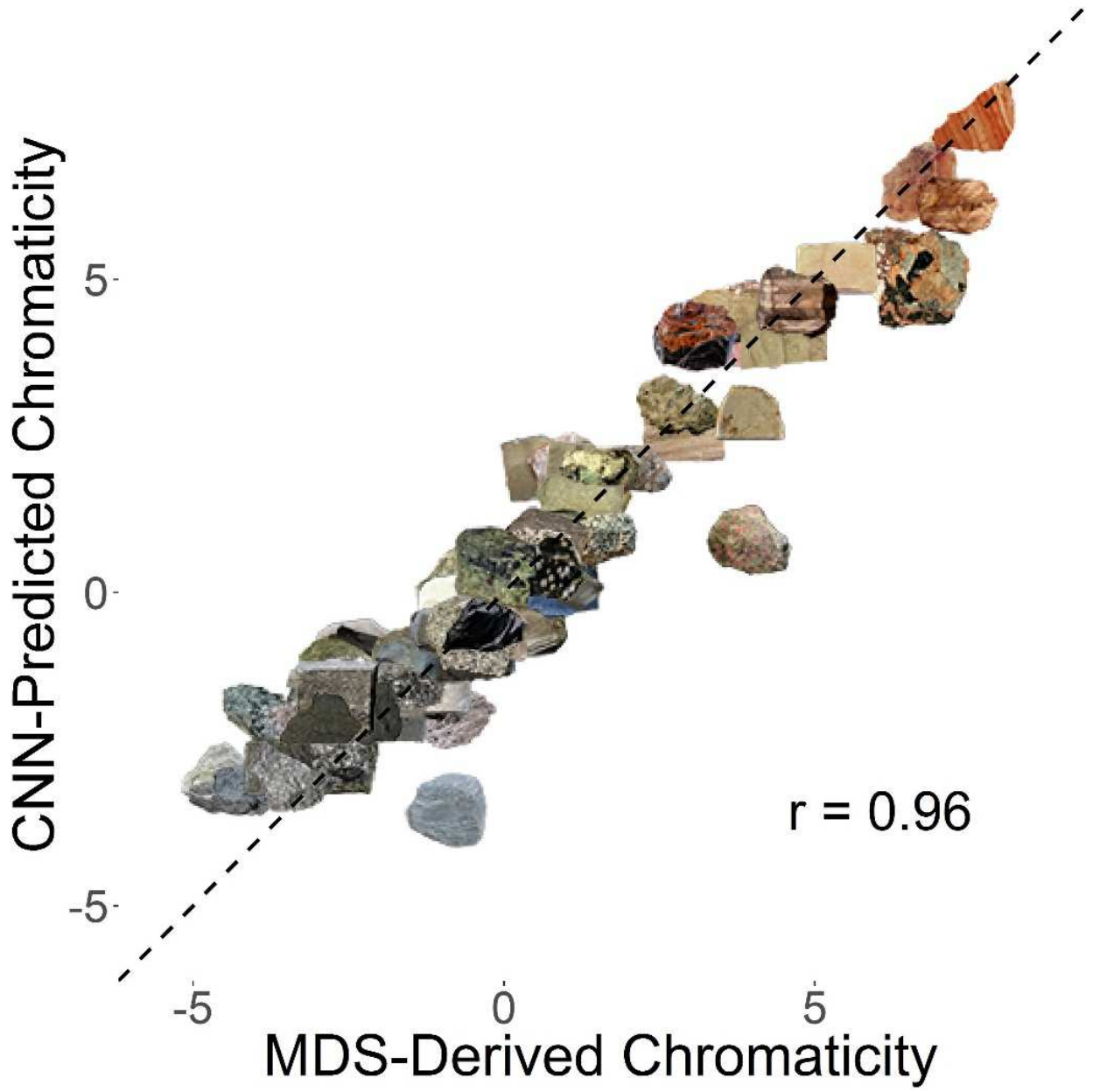

Figure 14: Scatterplot of Ensemble-predicted chromaticity against MDS-derived chromaticity for the test set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity.

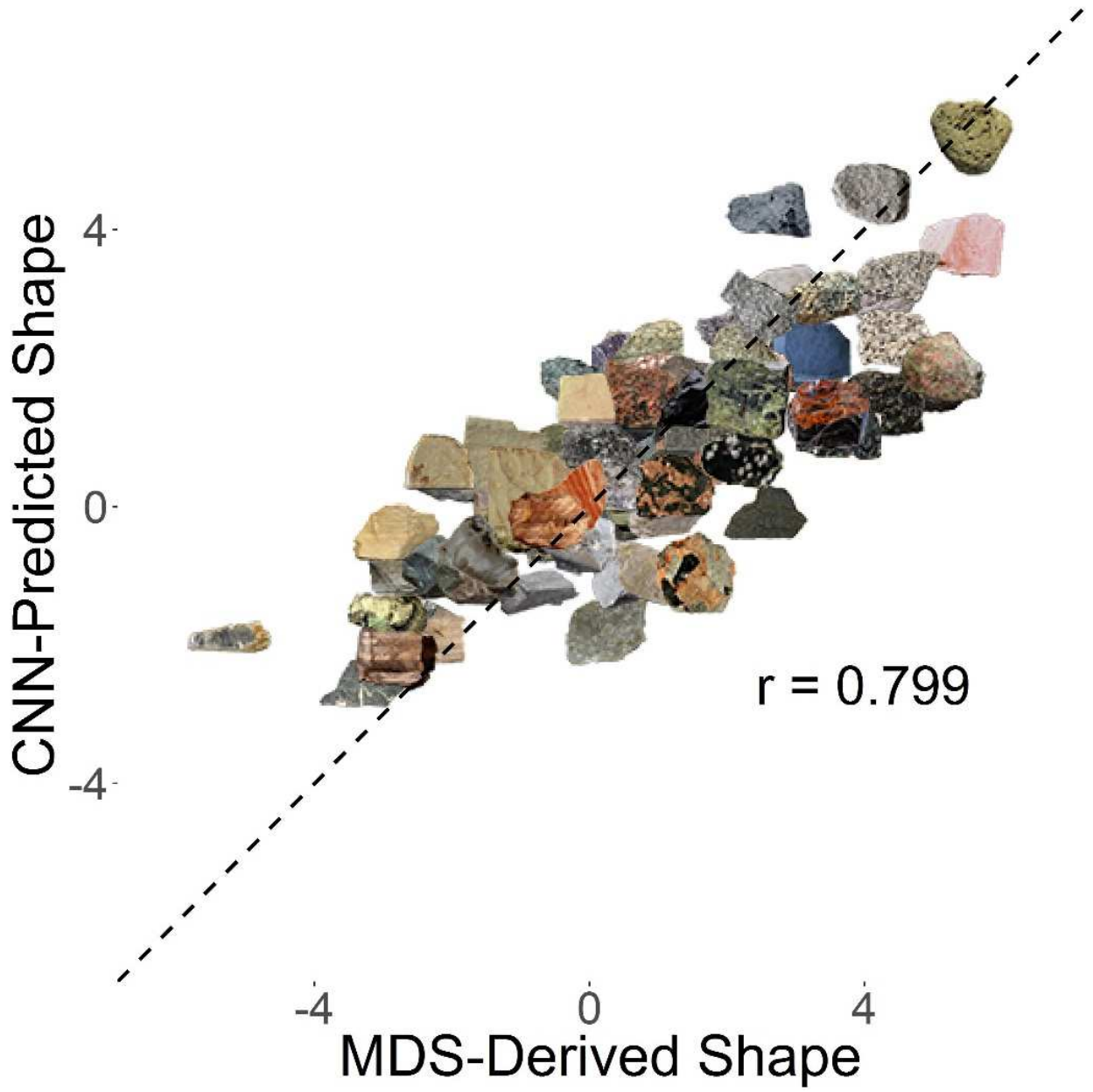

Figure 15: Scatterplot of Ensemble-predicted shape against MDS-derived shape for the test set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity. Note that "shape" is only a loose interpretation of this dimension.

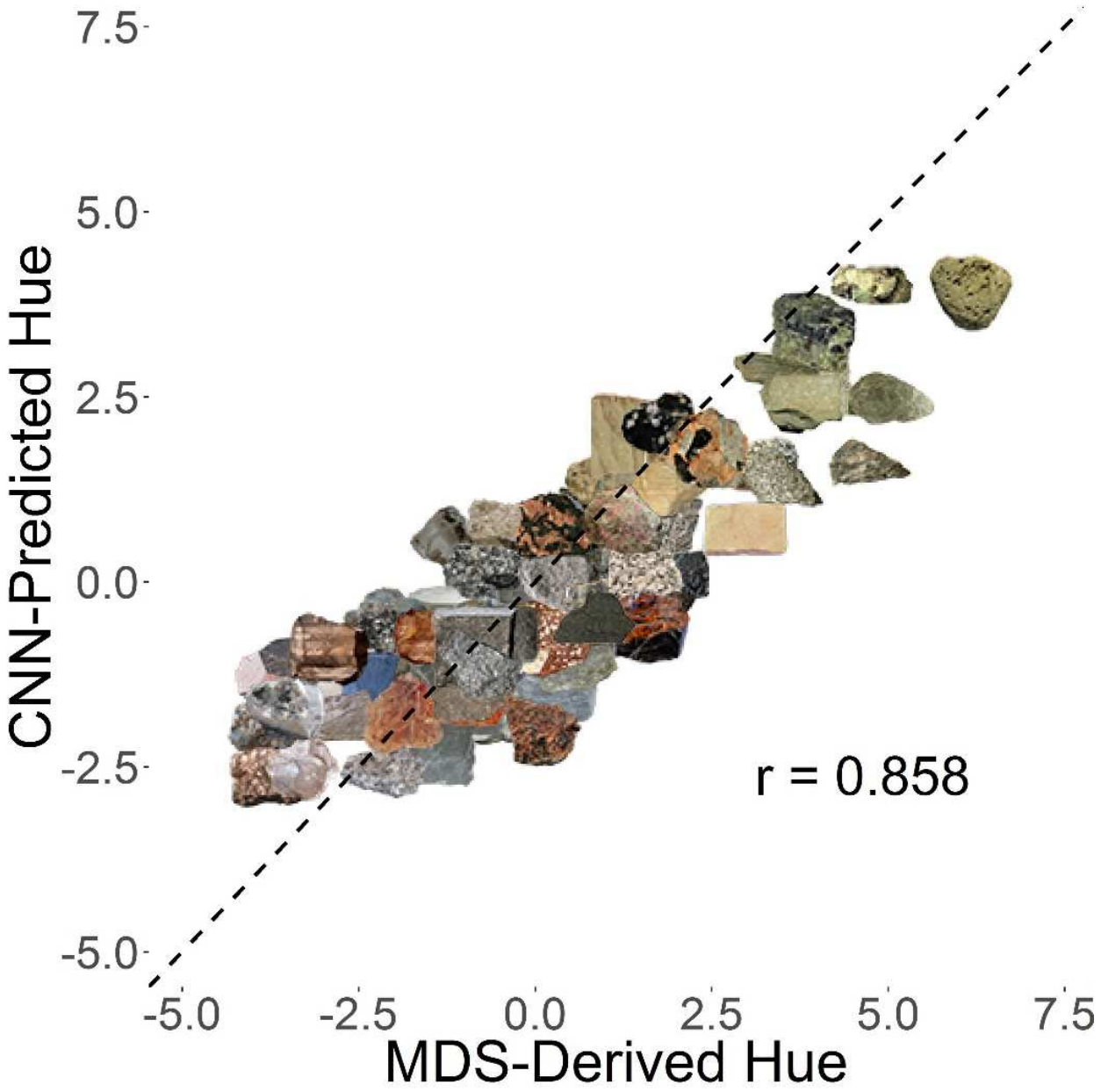

Figure 16: Scatterplot of Ensemble-predicted hue against MDS-derived hue for the test set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity. Note that "hue" is only a loose interpretation of this dimension.

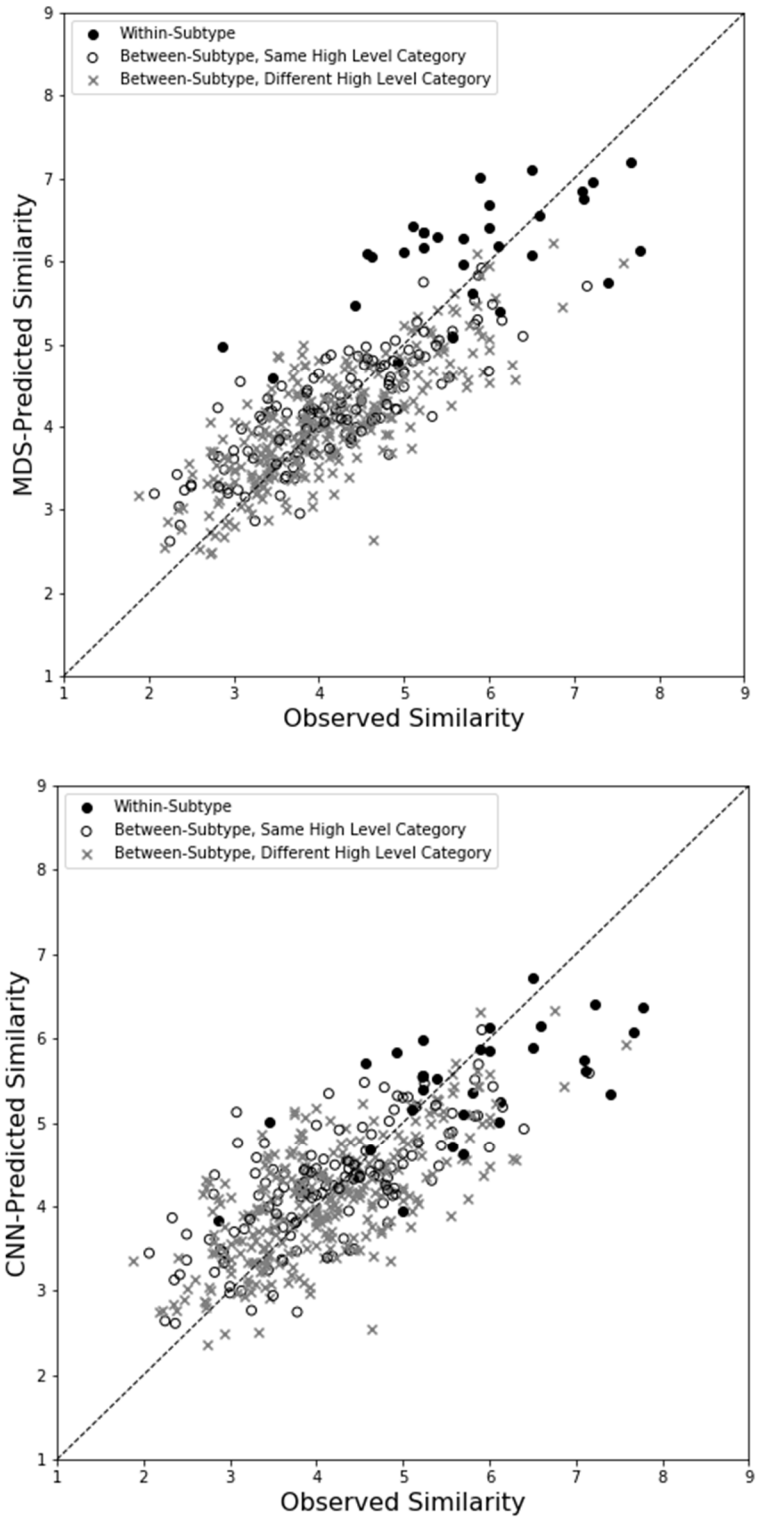

Figure 17: Scatterplot of observed similarity judgments against predicted similarities, collapsed across rock subtypes. Top panel: predictions derived from MDS-derived representations. Bottom panel: predictions derived from CNN-derived representations.

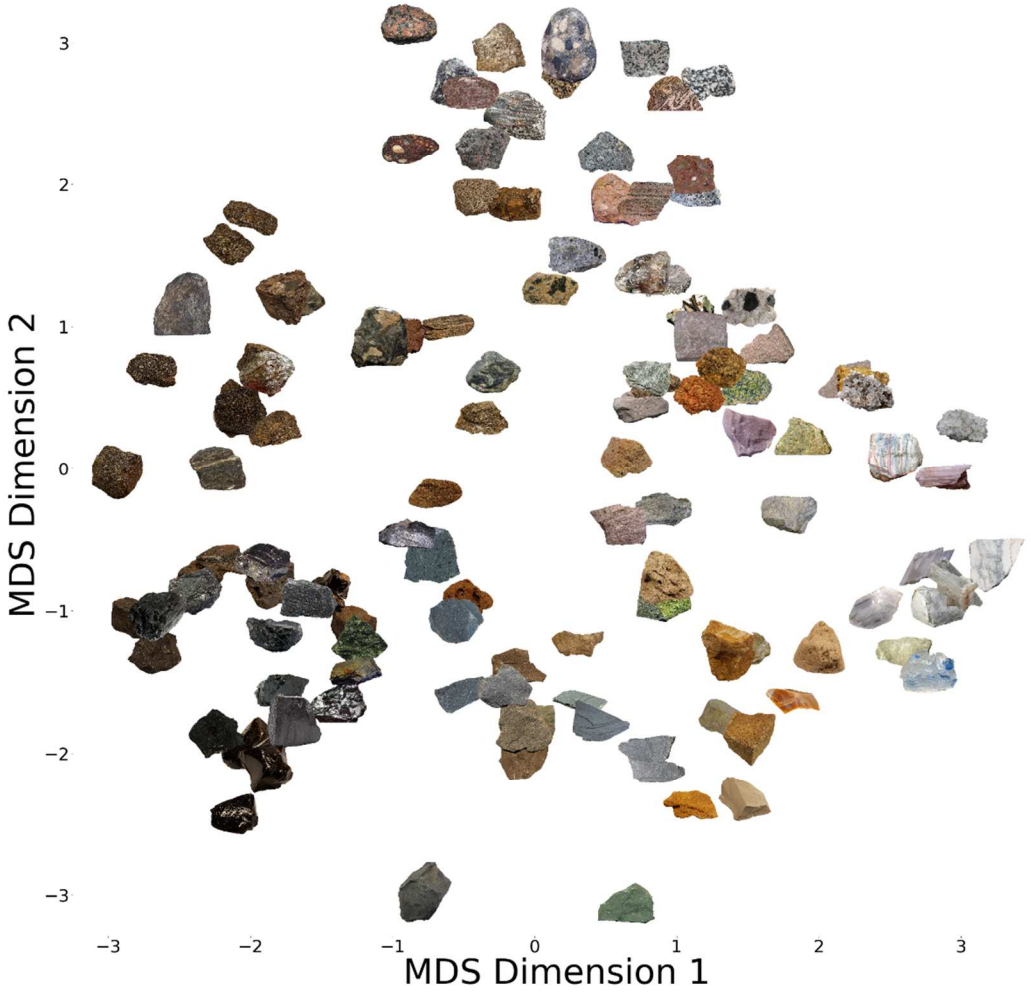

Figure 18: Plot of the first two rotated MDS dimensions. Dimension 1 can be interpreted as the rocks' lightness/darkness, and dimension 2 can be interpreted as the rocks' average grain size.

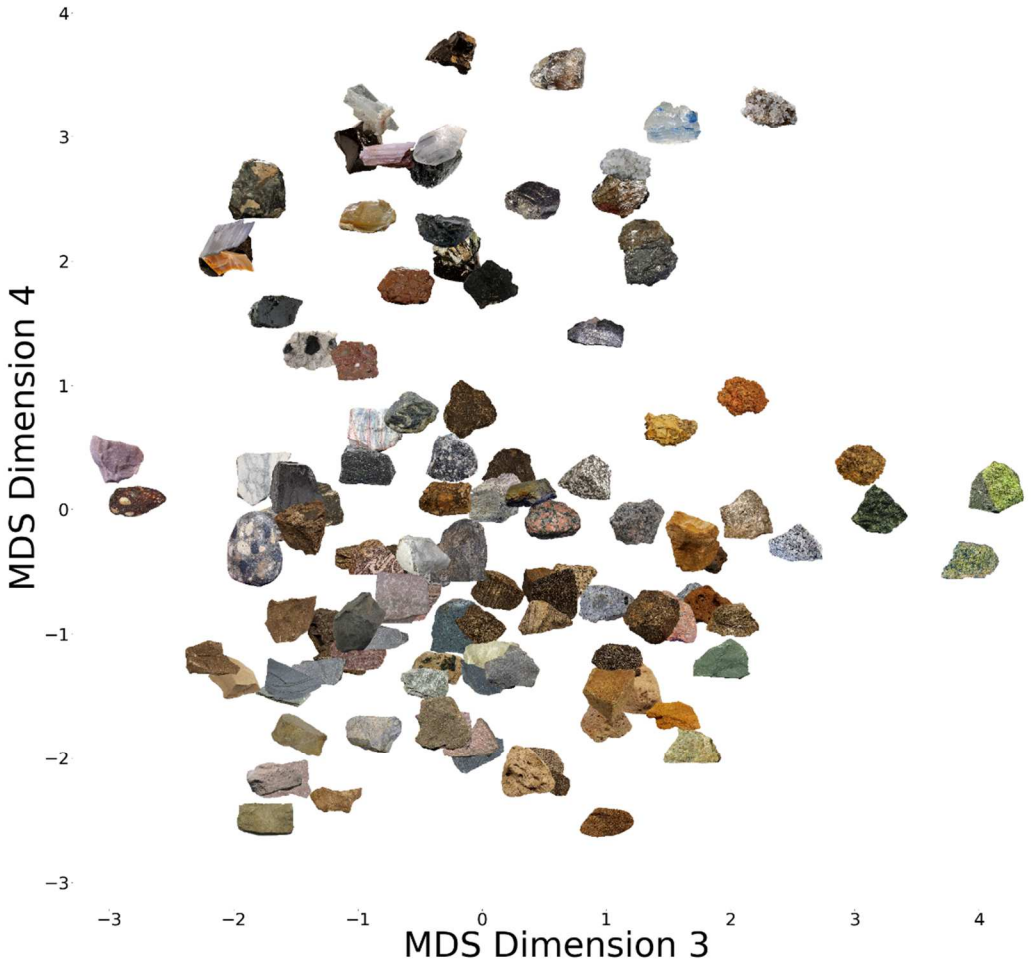

Figure 19: Plot of the third and fourth rotated MDS dimensions. Dimension 3 can be interpreted as the rocks' roughness (although the correlation with the direct roughness ratings was relatively low, so this interpretation should be taken with some caution), and dimension 4 can be interpreted as the rocks' shininess.

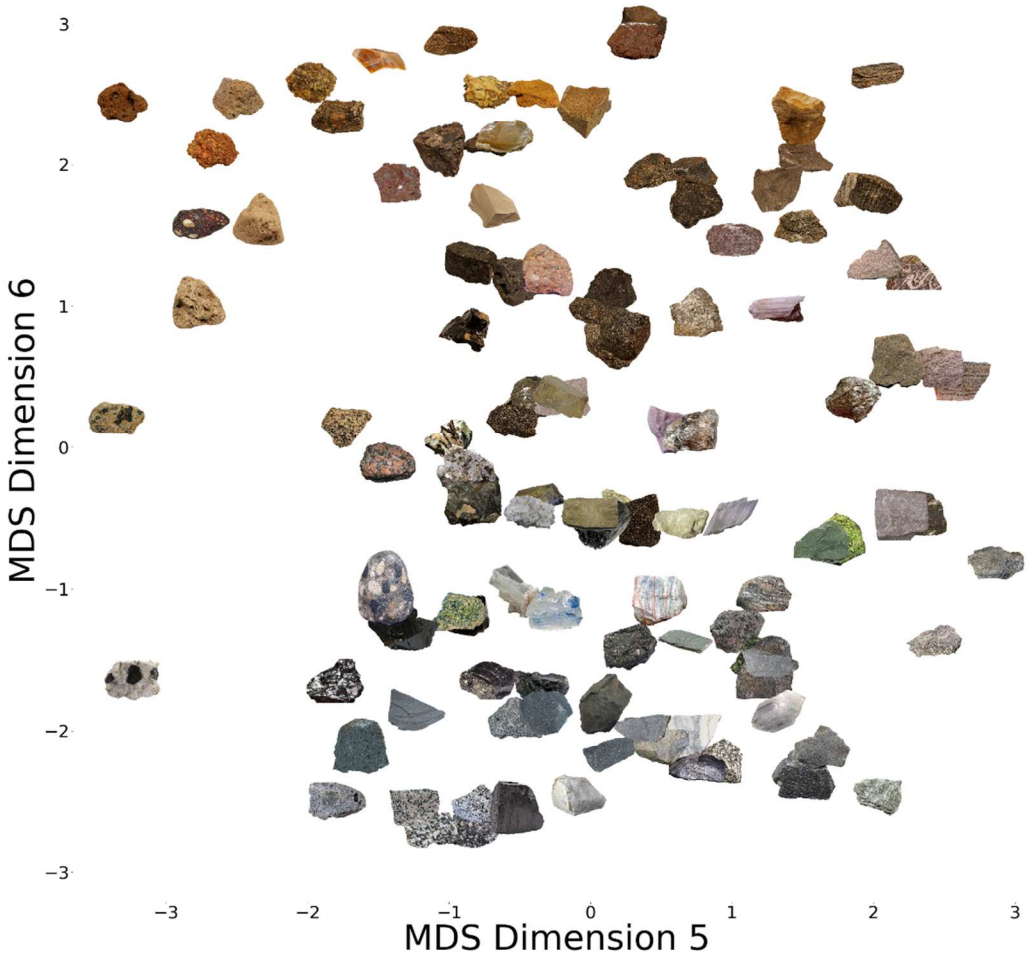

Figure 20: Plot of the fifth and sixth rotated MDS dimensions. Dimension 5 can be interpreted as the rocks' organization (the extent to which a rock has organized layers or stripes vs. fragments haphazardly glued together; note, though that the correlation with the direct organization ratings was relatively low, so this interpretation should be taken with some caution), and dimension 6 can be interpreted as the rocks' chromaticity (the extent to which a rock's color is saturated/warm or desaturated/cool).

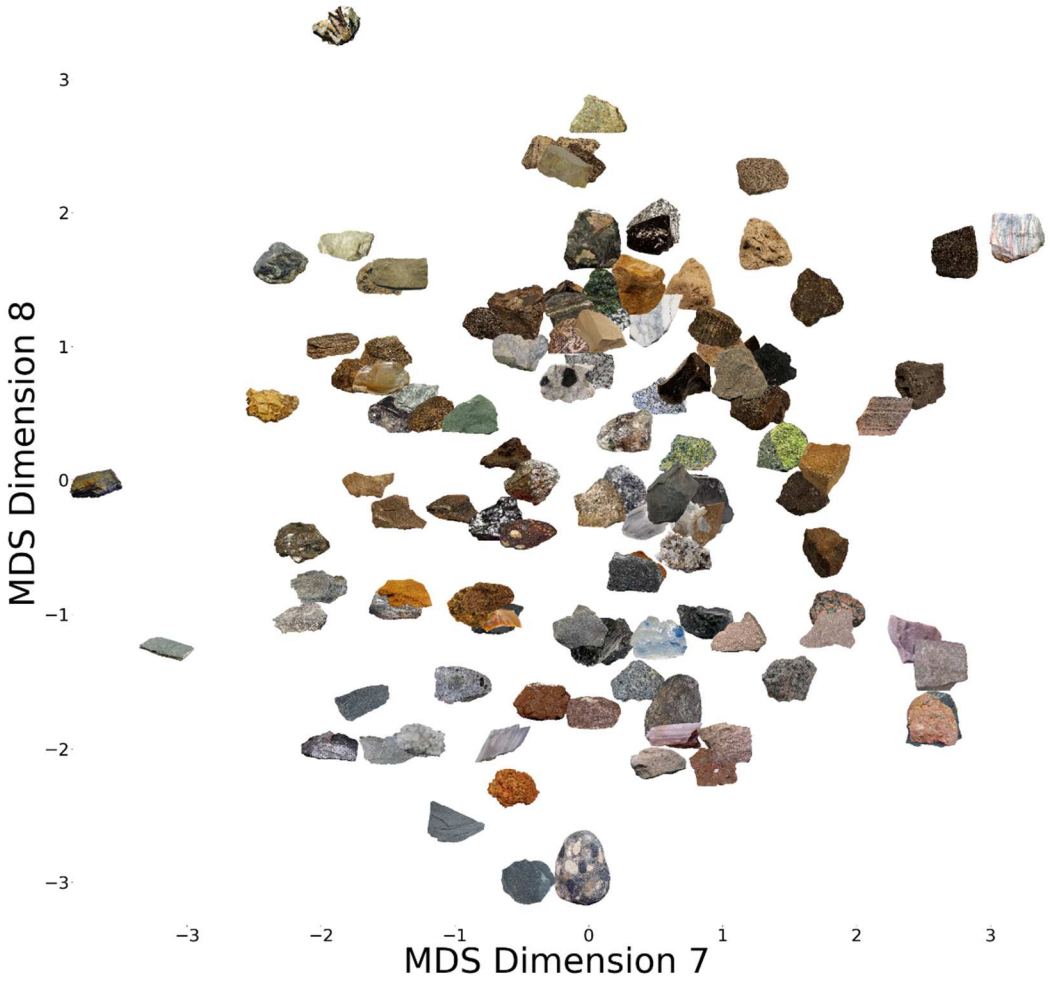

Figure 21: Plot of the seventh and eight rotated MDS solution. While these dimensions do not have clear interpretations, dimension 7 seems to have some correspondence with the rocks' shape (rocks on the left side of the space tend to be flat, while rocks on the right tend to have more volume), and dimension 8 seems to have some correspondence with the rocks' hue (rocks on the top of the space tend to be yellow/green, while rocks on the bottom tend to be red/violet).

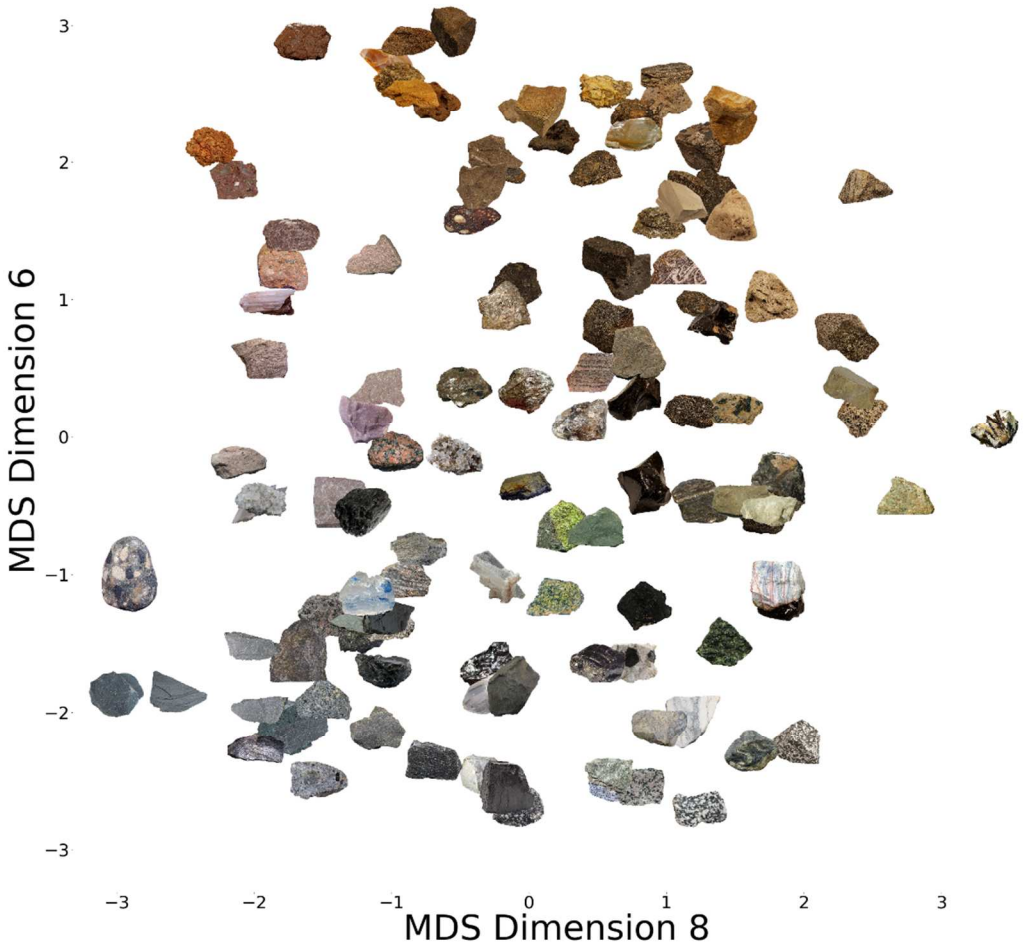

Figure 22: Plot of the eighth and sixth dimensions from the rotated MDS solution. Plotting these dimensions together forms a loose color circle. Starting from the top and moving clockwise, the color of the rocks shifts from red, to orange, to yellow, to green, to blue, to violet, and then finally back to red. Achromatic rocks are located to the lower-left of the plot.

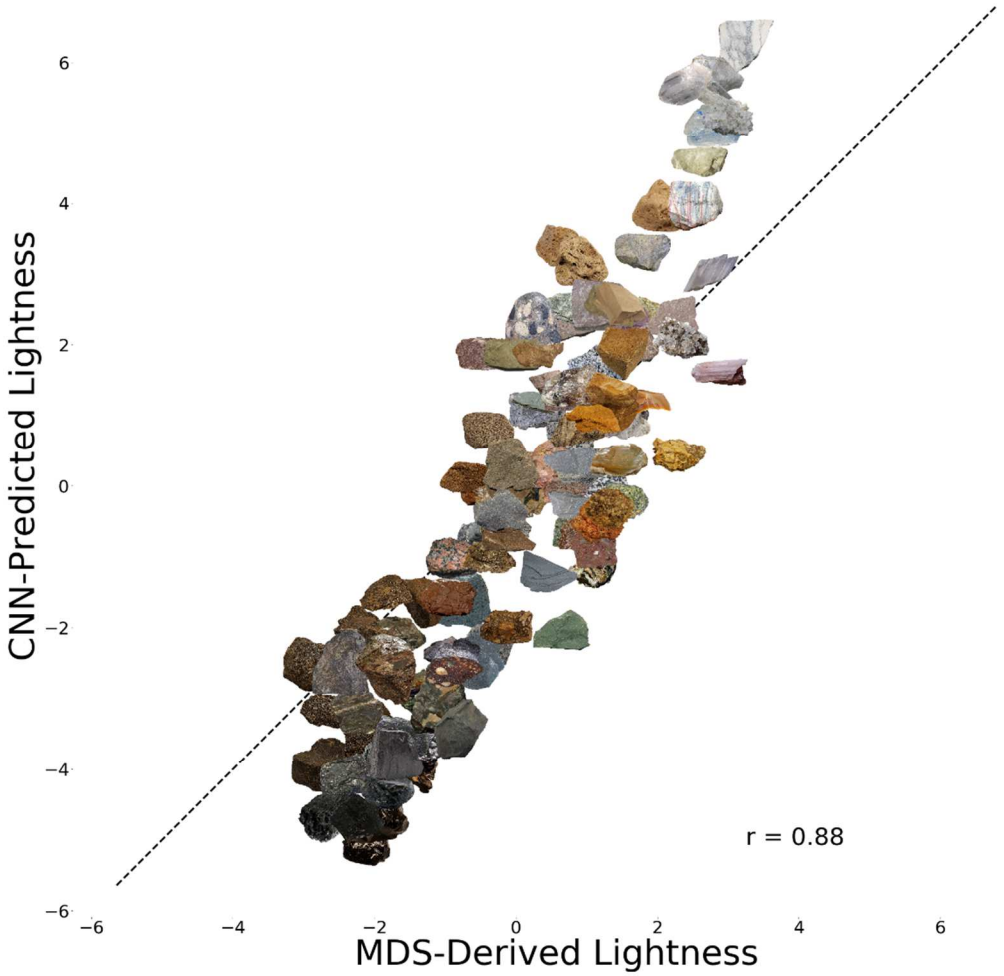

Figure 23: Scatterplot of Ensemble-predicted lightness against MDS-derived lightness for the 120 rocks set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity.

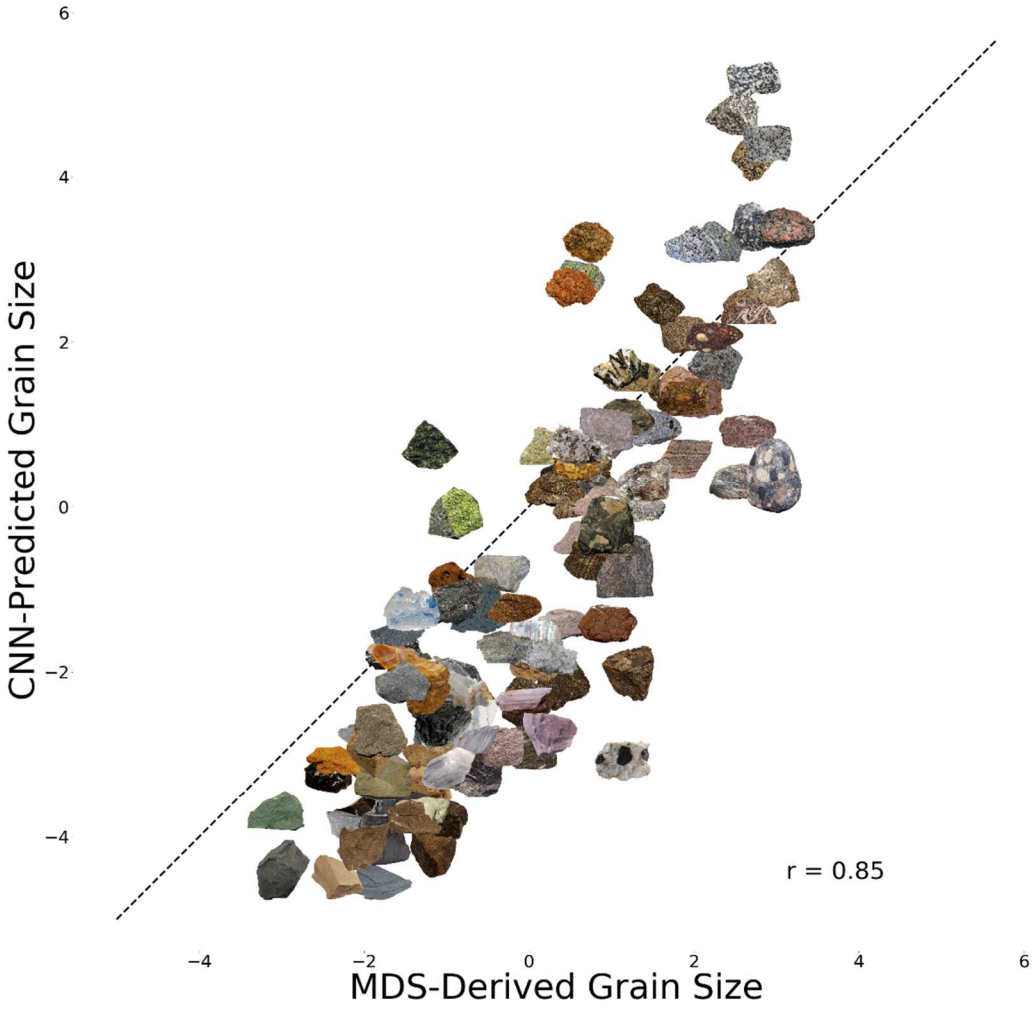

Figure 24: Scatterplot of Ensemble-predicted grain size against MDS-derived grain size for the 120 rocks set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity.

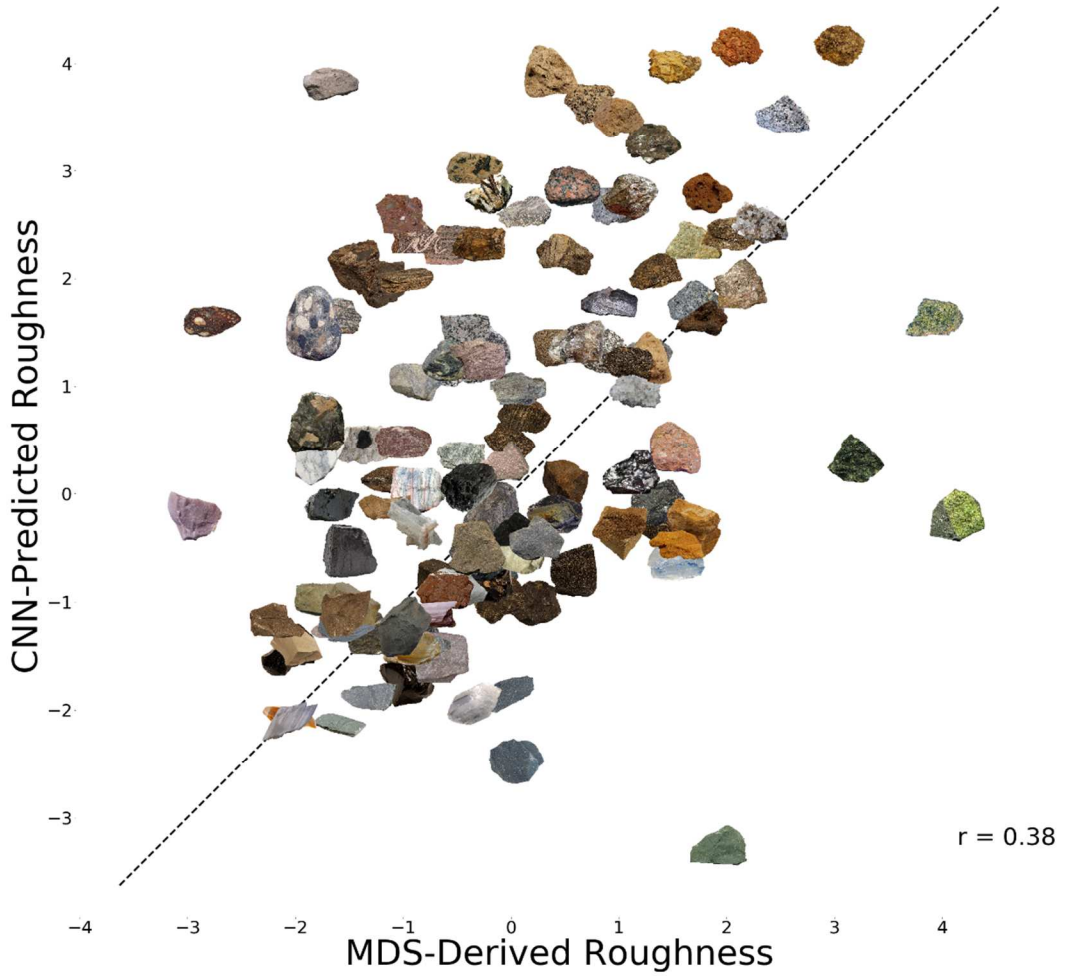

Figure 25: Scatterplot of Ensemble-predicted roughness against MDS-derived roughness for the 120 rocks set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity. Note that the "roughness" interpretation should be taken with caution as there was a low correlation between this MDS dimension and the direct roughness ratings.

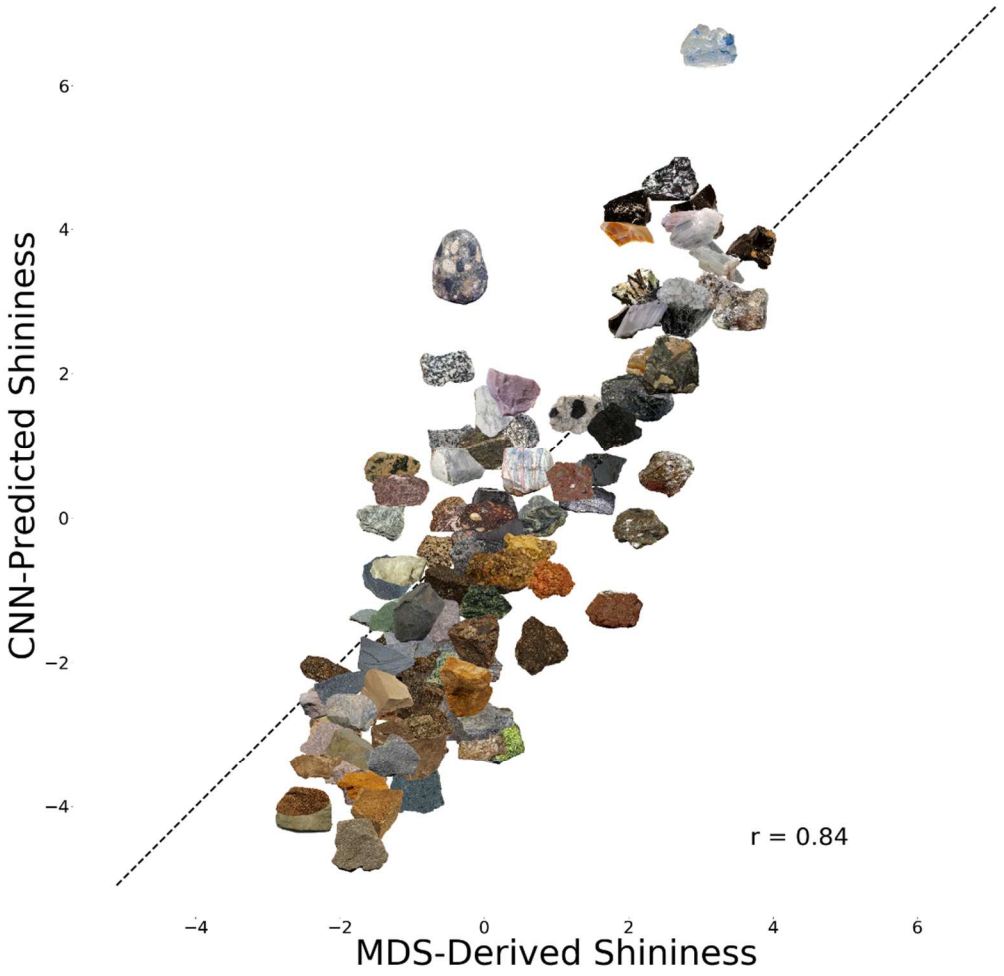

Figure 26: Scatterplot of Ensemble-predicted shininess against MDS-derived shininess for the 120 rocks set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity.

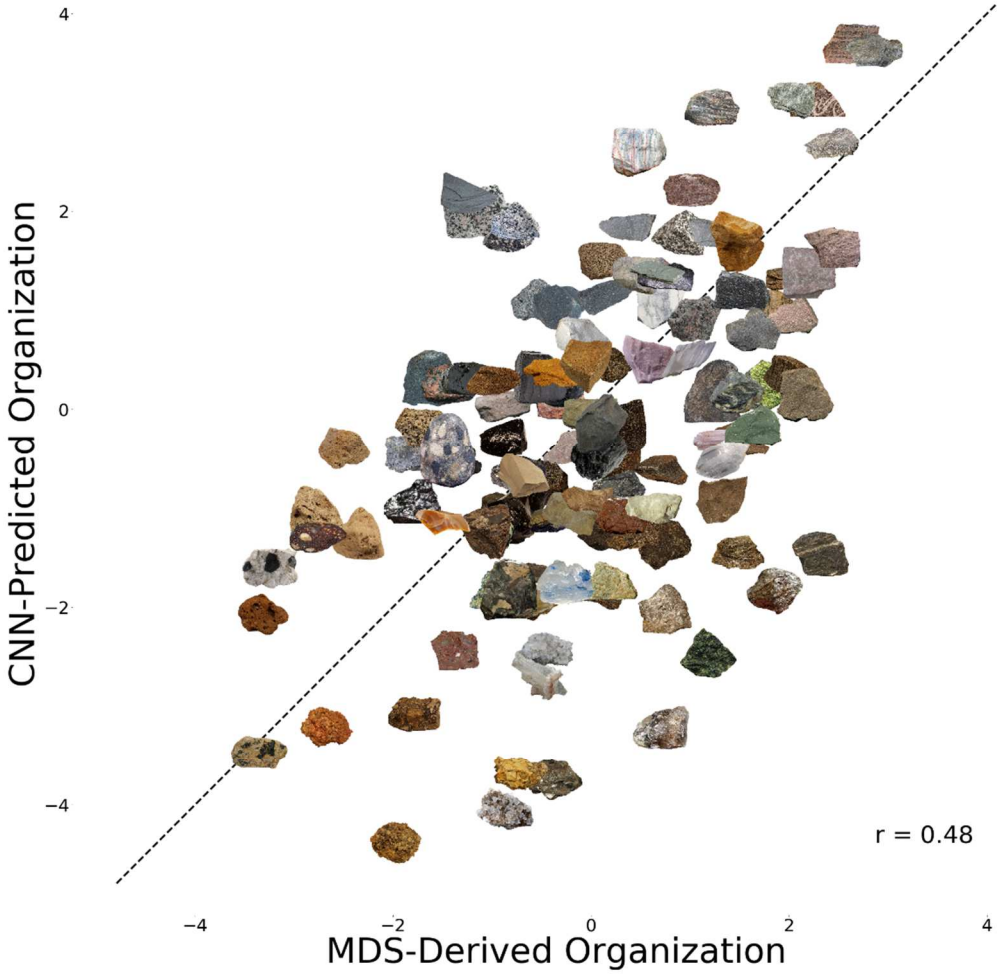

Figure 27: Scatterplot of Ensemble-predicted organization against MDS-derived organization for the 120 rocks set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity. Note that the "organization" interpretation should be taken with caution as there was a low correlation between this MDS dimension and the direct organization ratings.

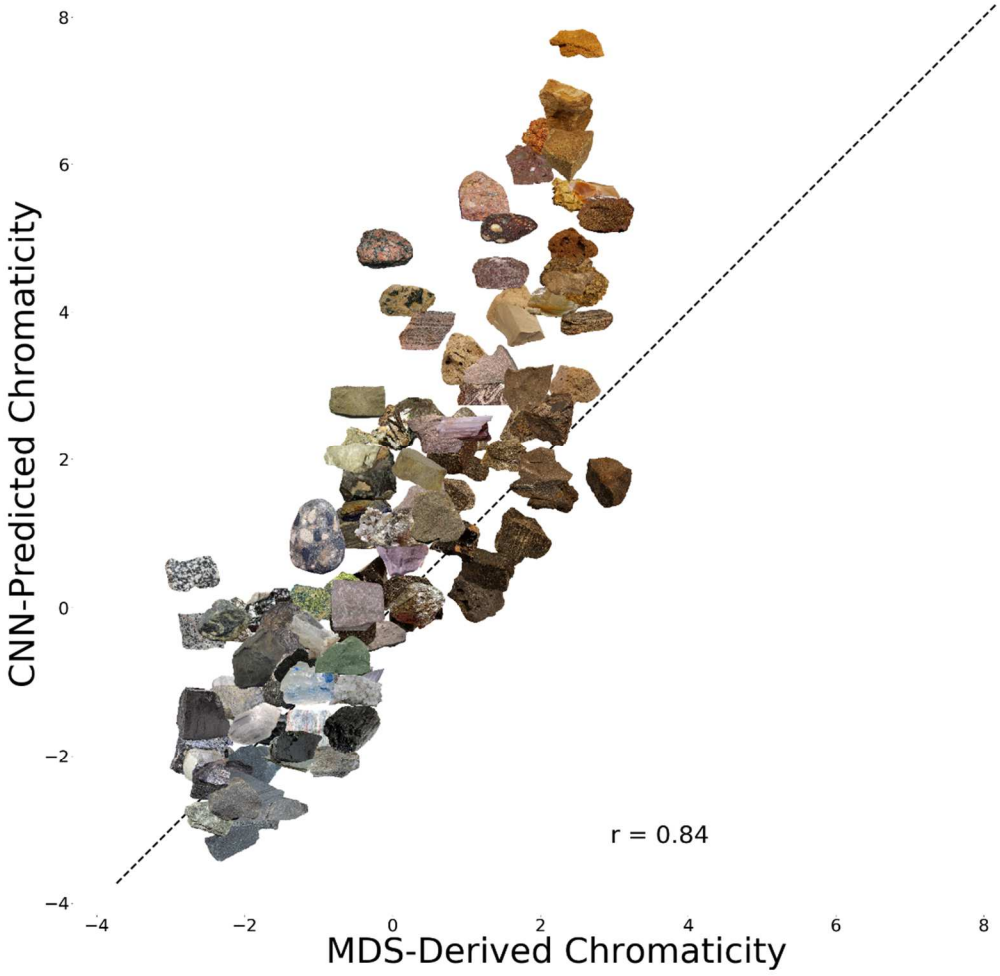

Figure 28: Scatterplot of Ensemble-predicted chromaticity against MDS-derived chromaticity for the 120 rocks set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity.

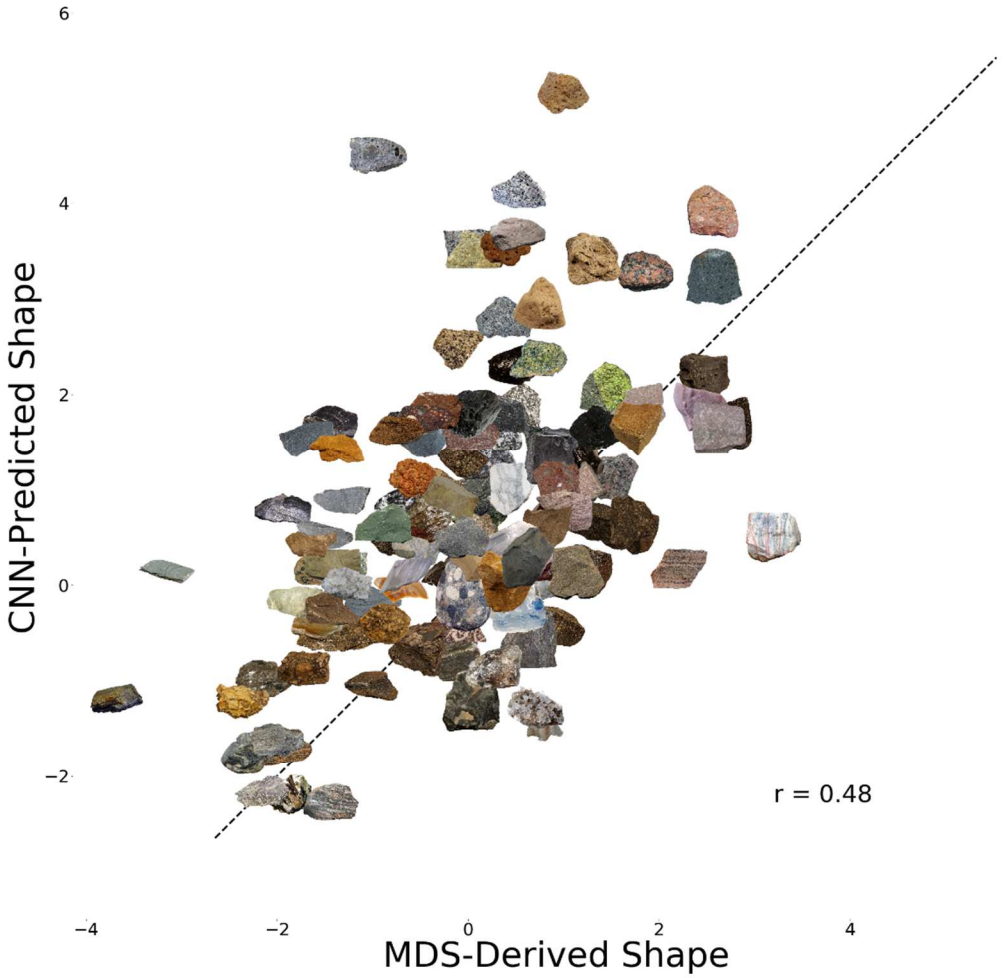

Figure 29: Scatterplot of Ensemble-predicted shape against MDS-derived shape for the 120 rocks set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity. Note that shape is only a loose interpretation of this dimension.

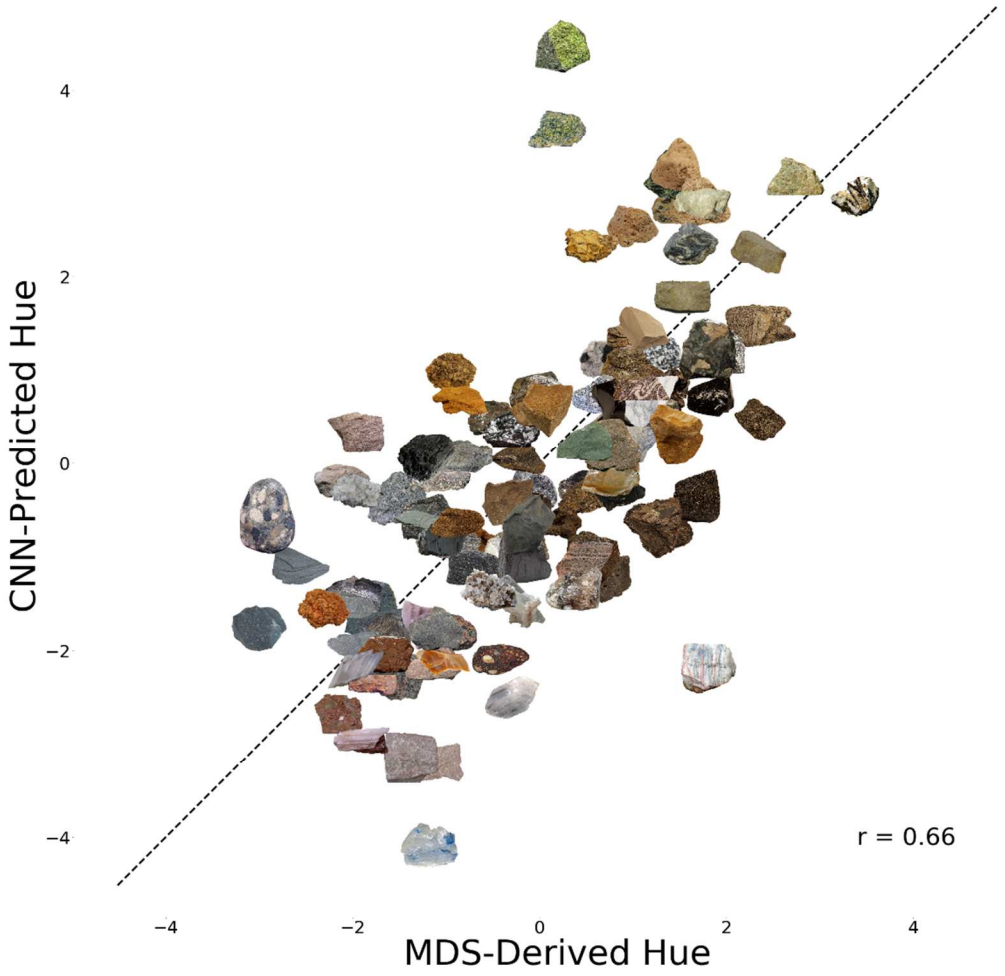

Figure 30: Scatterplot of Ensemble-predicted hue against MDS-derived hue for the 120 rocks set. The r value indicates the Pearson correlation coefficient, and the dashed line represents unity. Note that hue is only a loose interpretation of this dimension.

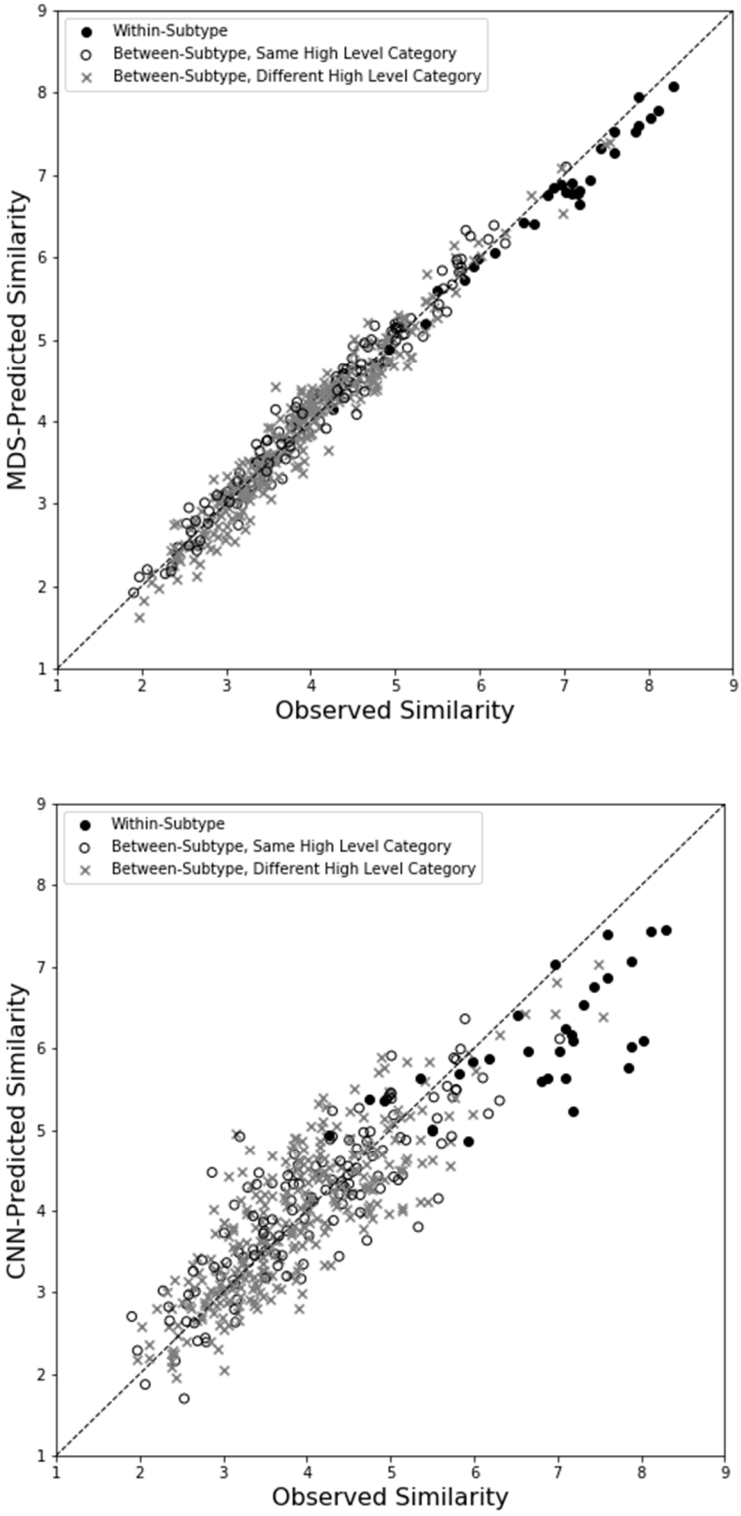

Figure 31: Scatterplot of observed similarity judgments against predicted similarities, collapsed across rock subtypes. Top panel: predictions derived from MDS-derived representations. Bottom panel: predictions derived from CNN-derived representations.

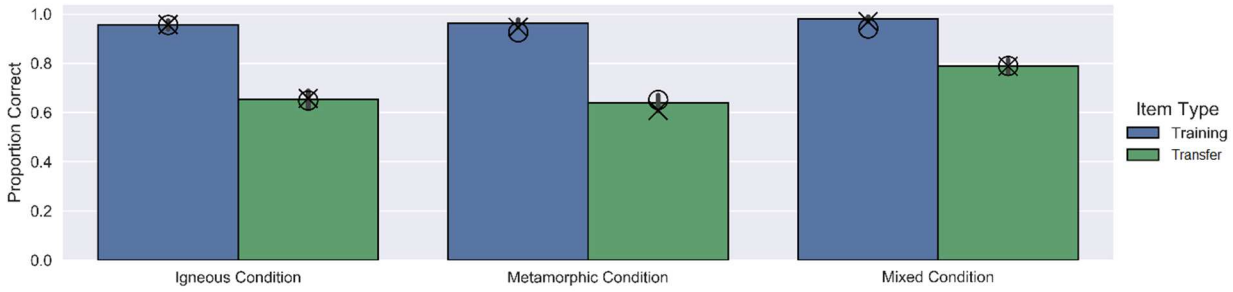

Figure 32: Mean proportion correct in the test phase as a function of condition and item type (training or transfer). Bar heights indicated observed data, error bars indicate 95% confidence intervals, circles indicate GCM predictions using CNN representations, and crosses indicate GCM predictions using MDS representations.

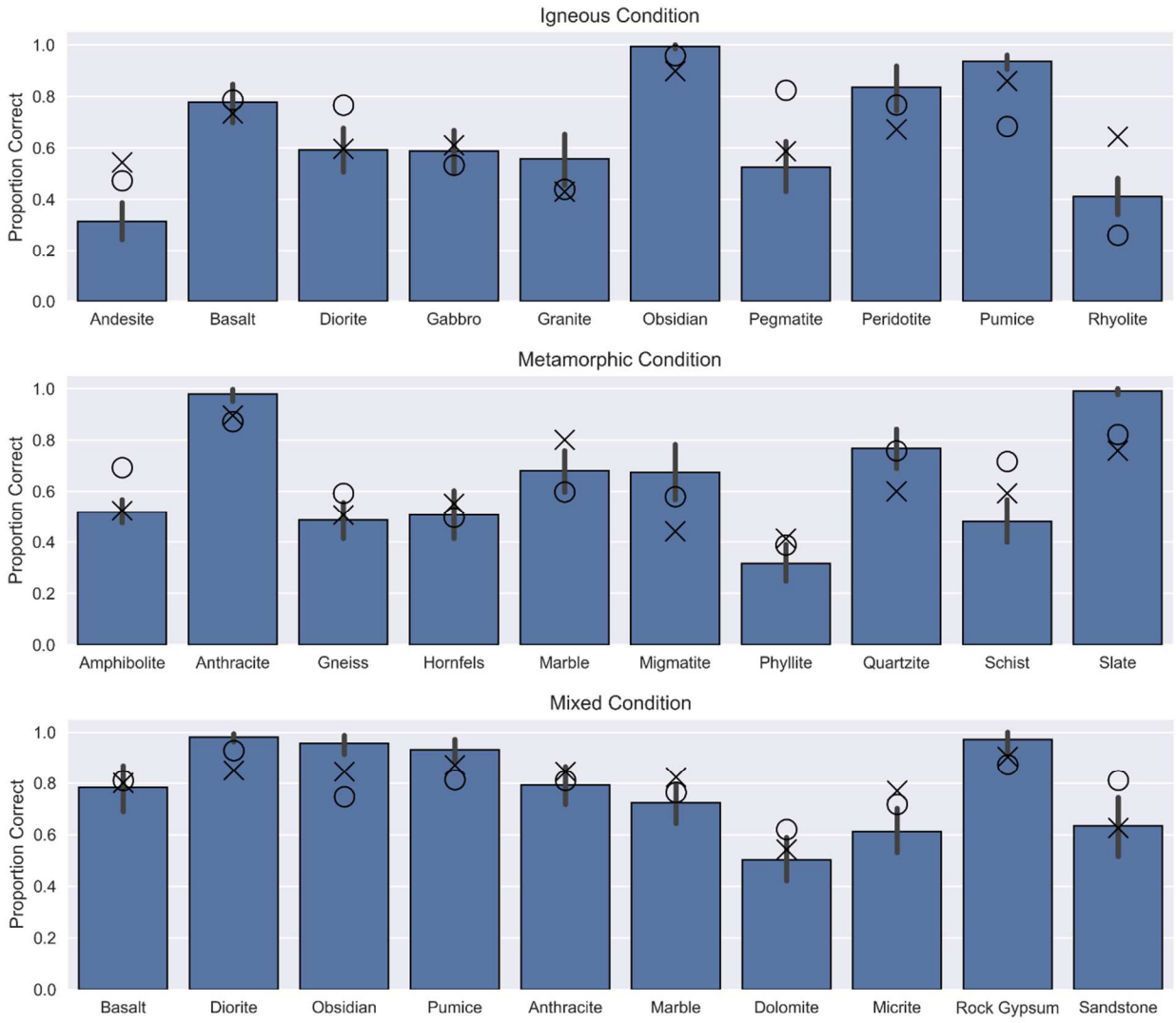

Figure 33: Mean proportion correct for transfer items as a function of condition and category of rock. Bar heights indicated observed data, error bars indicate 95% confidence intervals, circles indicate GCM predictions using CNN representations, and crosses indicate GCM predictions using MDS representations.

	Andesite			Basalt			Diorite			Gabbro			Granite			Obsidian			Pegmatite			Peridotite			Pumice			Rhyolite		
	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS
Andesite 1	95	99	97	1	0	1	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0
Andesite 2	88	92	91	0	1	0	2	2	3	1	1	0	5	0	3	0	0	0	1	4	2	1	0	1	1	0	0	1	0	0
Andesite 3	25	16	44	1	2	0	48	48	22	1	2	2	22	21	25	0	0	0	1	7	4	1	2	1	0	0	0	1	2	1
Andesite 4	69	47	51	1	1	0	8	23	13	3	0	1	14	18	27	0	0	0	1	7	2	2	2	3	1	0	1	2	2	4
Basalt 1	0	0	0	83	94	93	0	1	0	14	5	3	0	0	0	2	0	3	0	0	0	0	0	0	0	0	0	0	0	0
Basalt 2	0	1	0	92	95	97	0	1	0	4	2	1	0	0	0	0	0	1	1	0	0	0	0	0	0	2	0	0	0	0
Basalt 3	2	3	4	85	80	77	0	1	0	8	7	10	1	1	1	0	0	3	2	0	0	0	1	1	0	8	2	1	1	1
Basalt 4	5	2	1	73	76	80	0	1	0	16	17	8	0	1	0	0	0	2	0	0	0	1	0	3	1	3	4	3	0	1
Diorite 1	0	0	1	0	0	0	92	90	91	0	2	1	5	7	3	0	0	0	1	0	3	1	1	1	0	0	1	0	0	0
Diorite 2	0	0	2	0	0	0	98	96	89	0	1	0	2	4	7	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
Diorite 3	6	2	2	1	2	1	55	43	51	3	8	15	6	29	6	0	1	0	1	2	3	18	5	9	6	5	9	6	2	5
Diorite 4	0	2	3	0	2	0	93	75	69	0	1	2	6	18	19	0	0	0	0	2	2	0	0	4	0	0	0	0	0	1
Gabbro 1	1	0	0	3	2	1	1	0	1	86	96	97	4	2	0	0	0	0	4	0	1	0	0	0	0	0	0	1	0	0
Gabbro 2	1	0	0	18	13	5	0	0	0	79	87	92	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
Gabbro 3	4	1	1	4	7	5	6	8	3	64	80	86	8	2	2	0	1	0	2	0	2	5	0	1	0	0	1	7	0	0
Gabbro 4	1	2	1	59	57	62	0	0	0	38	37	31	0	1	0	2	2	5	0	0	0	0	0	0	0	0	1	0	0	0
Granite 1	0	0	1	0	0	0	1	0	2	1	2	1	97	97	95	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1
Granite 2	1	1	3	0	0	0	2	7	8	0	2	0	97	90	86	0	0	0	0	0	0	0	0	2	0	0	0	0	0	1
Granite 3	1	2	5	0	0	0	34	32	61	0	1	0	64	64	31	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1
Granite 4	5	10	5	4	2	1	3	10	5	8	2	2	74	48	74	0	0	0	1	18	1	0	4	4	0	1	1	3	6	8
Obsidian 1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	100	100	99	0	0	0	0	0	0	0	0	0	0	0	0
Obsidian 2	0	0	0	1	0	2	0	0	0	0	0	1	0	0	0	99	100	97	0	0	0	0	0	0	0	0	0	0	0	0
Obsidian 3	0	0	0	1	1	2	0	0	0	0	0	2	0	0	0	99	99	92	0	0	5	0	0	0	0	0	0	0	0	0
Obsidian 4	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	99	99	98	0	0	0	0	0	0	0	1	0	0	0	0
Pegmatite 1	0	1	1	0	1	0	2	4	2	0	1	0	1	0	0	0	0	0	95	88	94	1	5	1	0	1	1	0	0	0
Pegmatite 2	0	0	0	1	0	0	1	0	0	3	0	0	2	0	0	0	0	0	92	100	99	0	0	0	0	0	0	0	0	0
Pegmatite 3	0	4	2	0	1	1	9	6	8	0	1	1	5	6	3	0	5	1	83	56	77	1	16	5	0	0	2	0	5	0
Pegmatite 4	4	7	8	9	0	0	2	13	19	4	3	0	3	20	12	5	1	1	58	49	51	5	2	4	8	1	4	2	3	1
Peridotite 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	99	99	0	0	0	0	1	0
Peridotite 2	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	4	0	90	93	97	8	1	1	0	2	0
Peridotite 3	2	0	3	0	0	0	6	1	3	2	1	0	2	4	14	0	0	0	4	2	1	83	85	72	2	2	4	0	5	2
Peridotite 4	1	2	2	6	2	9	1	1	0	6	4	3	1	1	1	1	0	6	4	4	1	78	82	75	0	0	1	0	3	0
Pumice 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2	99	99	97	0	1	0
Pumice 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	0	1	98	100	96	0	0	1
Pumice 3	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	1	5	97	98	91	0	1	1
Pumice 4	2	1	2	4	7	7	2	0	1	9	2	1	7	0	1	0	0	1	6	0	3	4	1	2	59	90	83	8	0	1
Rhyolite 1	0	1	1	0	0	0	0	0	1	1	0	0	1	0	3	0	0	0	0	0	1	0	0	1	0	1	1	98	99	93
Rhyolite 2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	0	2	99	99	94
Rhyolite 3	5	26	4	1	2	1	5	12	20	14	5	6	10	4	24	0	1	0	4	16	5	4	10	4	1	0	2	56	23	33
Rhyolite 4	20	20	8	1	2	1	3	1	3	1	1	1	6	3	2	0	0	0	1	2	3	8	2	7	36	9	11	23	59	63

Figure 34: Igneous condition confusion matrix. Each row represents one rock and each column represents one category. Within cell i,j, the middle value indicates the observed proportion of times rock i was categorized into j, while the left and right numbers indicate GCM+CNN and GCM+MDS predictions, respectively. Darker shading indicates higher values. For example, this table indicates that participants correctly categorized Andesite 3 16% of the time, while GCM+CNN predicted it would be categorized correctly 25% of the time, and GCM+MDS predicted it would be categorized correctly 44% of the time

	Amphibolite			Anthracite			Gneiss			Hornfels			Marble			Migmatite			Phyllite			Quartzite			Schist			Slate		
	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS
Amphibolite 1	79	98	95	1	0	0	1	0	1	4	0	1	0	0	0	1	0	0	5	0	0	1	0	0	5	2	3	4	0	0
Amphibolite 2	85	100	97	5	0	0	1	0	1	4	0	1	0	0	0	1	0	0	2	0	0	0	0	0	2	0	0	1	0	0
Amphibolite 3	44	8	37	2	2	1	1	3	3	9	17	16	0	1	1	1	0	3	28	22	9	1	1	3	6	45	16	8	1	11
Amphibolite 4	61	95	58	5	2	1	3	0	2	4	0	3	0	0	0	2	0	1	6	0	2	1	0	0	17	3	31	2	0	0
Anthracite 1	0	0	0	98	99	98	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
Anthracite 2	0	1	0	96	99	97	0	0	0	2	0	0	0	0	0	0	0	0	1	0	0	0	1	0	1	0	2	0	0	0
Anthracite 3	1	2	0	91	96	94	0	0	0	4	0	1	0	0	0	0	1	0	1	2	1	0	0	0	3	0	3	0	0	0
Anthracite 4	0	0	0	98	100	96	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2	0	0	0
Gneiss 1	1	1	1	0	0	0	93	97	95	0	1	1	0	0	0	1	0	2	1	1	0	2	0	0	2	1	1	1	0	0
Gneiss 2	0	0	1	0	0	0	91	96	91	0	0	0	0	0	0	3	2	4	2	1	0	2	1	3	1	0	0	0	0	0
Gneiss 3	3	0	1	0	1	1	62	26	48	1	5	10	0	0	1	8	9	18	15	40	8	2	1	2	7	9	8	3	10	2
Gneiss 4	12	6	26	1	1	1	42	72	44	3	6	13	0	0	0	4	10	4	14	1	1	1	0	2	18	3	8	4	0	0
Hornfels 1	1	2	1	7	1	1	0	1	1	84	95	94	0	0	0	0	0	1	7	1	1	0	0	0	1	1	1	1	0	0
Hornfels 2	2	0	0	2	0	0	0	1	0	85	94	92	0	0	0	0	0	0	1	1	1	0	1	0	3	2	0	7	2	5
Hornfels 3	8	1	0	2	0	0	1	1	0	67	53	48	0	0	1	1	1	0	2	9	4	1	1	1	3	5	1	15	30	43
Hornfels 4	20	13	10	2	3	5	4	2	3	38	48	55	0	0	1	5	5	9	8	15	3	3	2	4	13	12	7	7	0	2
Marble 1	0	0	0	0	0	0	0	1	1	0	0	0	99	99	97	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0
Marble 2	0	0	0	0	0	0	0	0	0	0	1	0	98	99	96	0	1	0	0	0	1	1	0	2	0	0	0	1	0	1
Marble 3	0	0	2	0	0	0	0	9	7	1	0	1	88	81	81	0	6	2	1	2	0	6	2	7	1	0	0	4	0	0
Marble 4	1	3	1	0	0	0	17	6	8	1	6	2	47	55	66	3	5	3	7	11	4	10	12	7	2	1	1	10	0	8
Migmatite 1	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	88	99	97	7	0	1	0	1	1	1	0	0	1	0	0
Migmatite 2	0	0	0	0	0	0	2	9	5	0	1	0	0	0	0	97	90	93	0	0	0	0	0	1	0	0	0	0	0	0
Migmatite 3	0	0	0	0	1	1	4	7	2	1	3	1	2	7	1	79	73	49	10	7	37	1	1	2	2	2	4	2	0	3
Migmatite 4	2	2	1	4	1	8	3	6	5	5	12	11	1	9	3	63	62	51	16	5	15	1	0	1	4	2	5	2	0	2
Phyllite 1	0	0	0	0	0	0	2	0	0	0	2	1	0	0	1	5	1	1	89	94	92	0	1	0	0	3	1	3	0	3
Phyllite 2	2	0	0	3	1	1	0	0	0	10	0	0	0	0	0	0	2	0	82	90	96	0	0	0	2	7	1	1	0	1
Phyllite 3	6	7	5	15	10	12	0	2	2	45	41	20	0	1	3	0	1	3	15	19	39	1	0	4	11	11	7	6	9	7
Phyllite 4	1	1	1	1	0	0	7	5	10	2	12	4	13	2	5	13	7	11	49	45	51	4	3	6	2	9	3	7	16	7
Quartzite 1	1	0	0	0	0	0	5	1	2	0	1	0	0	0	1	0	0	1	0	1	0	91	98	95	1	1	0	1	0	0
Quartzite 2	0	0	0	1	0	0	0	0	0	1	1	0	6	0	1	0	0	0	0	1	0	89	98	98	1	0	0	2	0	0
Quartzite 3	1	1	1	1	0	0	4	3	15	2	2	1	3	5	3	2	1	2	1	1	1	80	86	76	1	1	0	4	0	1
Quartzite 4	1	2	1	2	0	0	7	7	6	2	1	2	4	7	1	7	9	29	1	5	2	71	67	57	2	1	2	2	0	1
Schist 1	3	3	4	0	2	1	3	5	1	1	1	1	0	0	0	1	0	1	2	3	2	1	1	0	87	85	90	1	1	0
Schist 2	1	0	0	7	1	6	0	0	0	7	0	1	0	0	0	0	1	0	1	3	3	1	0	0	77	95	90	5	0	1
Schist 3	4	6	4	1	5	1	9	10	12	1	2	2	0	0	1	2	2	2	4	6	6	1	0	1	77	68	70	1	1	1
Schist 4	6	0	1	33	56	54	1	2	1	7	0	3	0	2	0	2	5	1	8	8	3	1	0	0	40	28	37	1	0	0
Slate 1	0	0	0	0	0	0	0	0	0	2	0	1	1	0	0	0	0	0	1	0	1	1	0	0	1	0	0	94	100	98
Slate 2	1	0	0	0	0	0	0	0	0	4	1	2	0	0	0	0	0	0	1	0	2	0	0	0	1	0	0	92	99	95
Slate 3	1	0	0	0	0	0	1	0	0	6	0	4	1	0	1	0	0	0	1	0	1	1	0	0	1	0	0	86	100	92
Slate 4	1	0	0	0	0	0	0	0	0	5	1	15	0	0	1	0	0	0	2	0	1	0	0	1	2	1	0	89	98	81

Figure 35: Metamorphic condition confusion matrix. Each row represents one rock and each column represents one category. Within cell i,j, the middle value indicates the observed proportion of times rock i was categorized into j, while the left and right numbers indicate GCM+CNN and GCM+MDS predictions, respectively. Darker shading indicates higher values. For example, this table indicates that participants correctly categorized Amphibolite 3 8% of the time, while GCM+CNN predicted it would be categorized correctly 44% of the time, and GCM+MDS predicted it would be categorized correctly 37% of the time.

	Basalt			Diorite			Obsidian			Pumice			Anthracite			Marble			Dolomite			Micrite			Rock Gypsum			Sandstone		
	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS	CNN	OBS	MDS
Basalt 1	94	99	95	0	0	0	2	0	3	0	0	0	3	1	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0
Basalt 2	97	98	96	0	0	0	0	0	1	0	2	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
Basalt 3	92	84	83	1	0	1	0	0	4	3	15	5	1	1	1	0	0	0	1	1	3	1	1	2	0	0	0	0	0	1
Basalt 4	72	73	80	0	2	0	0	1	2	2	13	4	1	0	1	0	0	0	6	8	6	17	2	3	0	1	0	1	0	3
Diorite 1	0	0	0	100	99	99	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Diorite 2	0	0	0	100	99	99	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Diorite 3	0	0	1	92	97	87	0	1	0	5	1	7	0	0	0	0	0	0	2	1	2	0	0	1	0	0	0	0	0	1
Diorite 4	0	0	0	100	99	96	0	0	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0
Obsidian 1	0	0	1	0	0	0	90	99	98	0	0	0	9	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Obsidian 2	2	1	1	0	0	0	88	99	97	0	0	0	10	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Obsidian 3	2	0	1	0	0	0	84	92	85	0	1	0	14	7	12	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0
Obsidian 4	1	0	1	0	0	0	89	99	95	0	1	0	10	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pumice 1	0	1	0	0	0	0	0	1	0	99	99	97	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
Pumice 2	0	0	0	0	0	1	0	1	0	100	99	98	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Pumice 3	0	1	0	1	0	1	0	0	0	98	99	95	0	0	0	0	0	0	0	0	1	0	0	2	0	0	0	0	0	0
Pumice 4	2	12	6	3	0	1	0	0	1	94	87	88	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	1	2
Anthracite 1	1	0	0	0	0	0	22	2	4	0	0	0	77	98	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anthracite 2	4	1	2	0	1	0	7	0	3	0	0	0	88	98	95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Anthracite 3	6	4	4	0	0	0	16	16	21	0	0	0	77	80	72	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
Anthracite 4	1	0	1	0	0	0	29	21	16	0	0	0	68	79	81	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0
Marble 1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	96	100	96	0	0	0	0	0	1	3	0	1	0	0	0
Marble 2	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	94	99	96	0	0	1	0	0	1	5	1	1	0	0	0
Marble 3	0	0	0	2	0	4	0	0	1	0	0	3	1	1	0	89	97	83	1	0	1	0	2	1	7	0	4	0	0	2
Marble 4	3	0	0	4	1	7	0	0	0	5	0	3	1	1	0	58	48	58	23	40	21	4	9	6	1	2	1	1	0	2
Dolomite 1	1	1	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	94	95	99	2	4	0	0	0	0	1	0	0
Dolomite 2	1	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	0	85	98	97	11	2	2	0	0	0	1	0	0
Dolomite 3	3	0	0	0	0	2	0	0	0	9	0	4	1	0	0	9	9	8	47	64	69	23	23	9	6	2	1	1	2	7
Dolomite 4	2	2	0	0	1	1	0	1	0	8	32	6	0	0	0	1	1	2	65	37	61	22	24	26	0	2	0	1	1	3
Micrite 1	1	0	0	0	0	0	0	0	0	3	0	3	0	0	0	0	0	1	9	2	3	86	97	92	0	1	0	2	0	1
Micrite 2	4	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	10	12	3	84	87	94	0	0	0	1	0	1
Micrite 3	2	1	1	0	0	0	0	0	0	10	1	2	0	0	0	0	0	0	26	32	21	56	64	65	0	1	0	6	2	9
Micrite 4	4	7	5	0	0	0	0	1	2	3	1	2	0	0	0	0	0	1	19	32	17	71	59	66	0	1	1	2	0	6
Rock Gypsum 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3	0	1	0	0	0	0	0	0	96	100	99	0	0	0
Rock Gypsum 2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	99	99	99	0	1	0
Rock Gypsum 3	1	0	0	1	0	0	4	0	0	0	1	0	3	0	0	2	1	1	0	0	0	0	0	0	89	97	98	0	1	0
Rock Gypsum 4	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	5	2	3	0	0	0	0	0	0	94	98	96	0	0	0
Sandstone 1	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	1	4	0	1	3	0	2	0	0	0	91	100	94
Sandstone 2	0	0	0	0	0	0	0	0	0	3	0	1	0	0	0	0	0	0	4	0	1	3	1	2	0	1	0	90	98	94
Sandstone 3	1	4	2	0	5	1	0	0	0	10	9	26	0	0	0	0	0	2	14	5	7	6	15	12	0	0	0	69	62	50
Sandstone 4	2	10	12	1	3	2	0	0	0	9	6	17	0	0	0	0	1	0	18	5	10	7	10	7	0	1	0	63	65	51

Figure 36: Mixed condition confusion matrix. Each row represents one rock and each column represents one category. Within cell *i,j,* the middle value indicates the observed proportion of times rock *i* was categorized into *j*, while the left and right numbers indicate GCM+CNN and GCM+MDS predictions, respectively. Darker shading indicates higher values. For example, this table indicates that participants correctly categorized Sandstone 3 62% of the time, while GCM+CNN predicted it would be categorized correctly 69% of the time, and GCM+MDS predicted it would be categorized correctly 50% of the time.

Craig A. Sanders

crasanders@gmail.com

Education & Research Appointments

2018 –	Postdoctoral Scholar in Psychological Sciences Vanderbilt University Advisors: Thomas Palmeri & Isabel Gauthier
2013 – 2018	Ph.D. in Psychological and Brain Sciences Indiana University, Bloomington Minor: Computer Science Advisor: Robert Nosofsky
2009 – 2013	B.S. in Brain, Behavior, and Cognitive Science University of Michigan, Ann Arbor Minors: Computer Science and Linguistics Advisor: Richard Lewis

Publications

Sanders, C.A. & Nosofsky, R.M. (in press). Using Deep-Learning Representations of Complex Natural Stimuli as Input to Psychological Models of Classification. *Proceedings of the 40th Annual Conference of the Cognitive Science Society*, Madison, WI.

Nosofsky, R.M., **Sanders, C.A.,** McDaniel, M. (2018). A formal psychological model of classification applied to natural-science category Learning. *Current Directions in Psychological Science*, 27(2), 129–135.

Nosofsky, R.M., **Sanders, C.A.**, McDaniel, M. (2018). Tests of an exemplar-memory model of classification in a high-dimensional natural-science category domain. *Journal of Experimental Psychology: General*, 147(3), 328-353.

Nosofsky, R.M., **Sanders, C.A.**, Meagher, B.J., & Douglas, B.J (2017). Toward the development of a feature-space representation for a complex natural category domain. *Behavior Research Methods,* 1-27.

Nosofsky, R.M., **Sanders, C.A.**, Gerdom, A., Douglas, B., & McDaniel, M. (2017). On learning natural science categories that violate the family-resemblance principle. *Psychological Science*, 28, 104-114.

Posters & Presentations

Sanders, C.A. & Nosofsky, R. (2017). Using Deep-Learning Representations of Complex Natural Stimuli as Input to Psychological Models of Classification. Cognitive Lunch, Indiana University, Bloomington.

Sanders, C.A., & Nosofsky, R. (2017). Using deep-learning to automatically extract psychological representations of complex stimuli. Poster presented at the annual meeting of the Psychonomic Society, Vancover, BC.

Sanders, C.A. (2017). Using deep-learning to automatically extract psychological representations of complex stimuli. Gray Matters, Indiana University, Bloomington.

Nosofsky, R.M & **Sanders, C.A.** (2016). Optimal Training Examples in Real-World Classification Learning. Fifty-eighth Annual Meeting of the Psychonomic Society, Boston, MA.

Nosofsky, R.M**, Sanders, C.A.**, & Meagher, B. (2016). High-dimensional category representations. Fifty-seventh Annual Meeting of the Psychonomic Society, Boston, MA.

Nosofsky, R.M, **Sanders, C.A.**, & Meagher, B. (2016). Enhancing learning of natural categories through guidance of formal models of human classification. Forty-ninth Annual Mathematical Psychology Society Meetings, New Brunswick, N.J.

Nosofsky, R.M, **Sanders, C.A.**, Gerdom, A., Miyatsu, T., & McDaniel, M. (2015). Teaching real-world categories at low and high levels of a hierarchy. Fifty-Sixth Annual Meeting of the Psychonomic Society, Chicago, IL.

Sanders, C.A., & Nosofsky, R. (2015). Category learning and education. Poster presented at 2015 IGERT Research Showcase, Indiana University, Bloomington.

Sanders, C.A., & Nosofsky, R. (2015). Models of category learning applied to education. Poster presented at 2014 Psychological and Brain Sciences Research Symposium, Indiana University, Bloomington.

Sanders, C.A., & Nosofsky, R. (2015). Models of category learning applied to education. Poster presented at 2014 IGERT Research Showcase, Indiana University, Bloomington.

Miyatsu, T., **Sanders, C.A.**, McDaniel, M., Nosofsky, R. (2014). Optimal Training Sets in Natural Category Learning. Poster presented at the annual meeting of the Psychonomic Society, Long Beach, CA.

Sanders, C.A., Lewis, R., & Shvartsman, M. (2013). A computational model of regressive eye movements in reading. Poster presented at 2013 Psychology Research Forum, University of Michigan, Ann Arbor.

Honors and Awards

2017	College of Arts and Sciences Fall Travel Award
2017	Data on the Mind Workshop attendee
2013-2015	NSF IGERT Traineeship in the Dynamics of Brain-Body-Environment Systems

Teaching Experience

Spring 2018	Teaching Assistant for Statistical Techniques
Summer 2017	Course Instructor for Foundations in Mathematics and Science: Programming
Spring 2016	Lab Instructor for Research Methods in Psychology
Spring 2016	Guest Lecturer for Social Media Mining

Technical Skills

Programming languages: Python, R, MATLAB, C++, SQL, Stan, Javascript

Machine learning frameworks: Tensorflow, scikit-learn, Keras

Other software: Microsoft Office, SPSS, Photoshop, git, Unix